

The international journal of science / 19 January 2023

nature

LINES OF DESCENT

Finland's isolated history offers fresh view of disease genetics

Facing the strain
How the body's first exposure to COVID can skew future immunity

Researchers at play
Turning experiments into games can boost behavioural science

Record warming
Highest temperatures in 1,000 years for Greenland's ice sheet

VISIT…

LANZAROTE Caliente.com

Nature.2023.01.21

[Sat, 21 Jan 2023]

- **This Week**
- **News in Focus**
- **Books & Arts**
- **Opinion**
- **Work**
- **Research**

This Week

- **How the periodic table survived a war to secure chemistry's future** [17 Jan 2023]
  • A century ago, the discovery of hafnium confirmed the validity of the periodic table — but only thanks to scientists who stood up for evidence at a time of global turmoil.
- **PhD training is no longer fit for purpose — it needs reform now** [18 Jan 2023]
  • If researchers are to meet society's expectations, their training and mentoring must escape the nineteenth century.
- **Preprint review should form part of PhD programmes and postdoc training** [17 Jan 2023]
  • Institutions owe it to young researchers to prepare them for careers outside academia. Preprint review is a perfect opportunity.

- EDITORIAL
- 17 January 2023

How the periodic table survived a war to secure chemistry’s future

A century ago, the discovery of hafnium confirmed the validity of the periodic table — but only thanks to scientists who stood up for evidence at a time of global turmoil.



Hafnium is a transition metal named after the Latin name for Copenhagen (Hafnia), where the element was discovered.Credit: Klaudia Kilman/Alamy

Hafnium isn’t a particularly remarkable element. It’s not your explosive sodium, shimmering mercury or stinky sulfur. It’s a greyish metal and is commonly used as a neutron absorber in the control rods of nuclear power

plants and nuclear submarines, and as an insulator in computer chips. But hafnium's discovery, which was reported in *Nature* a century ago this week[1], was of disproportionate importance. The element was identified by two scientists working in Copenhagen: Dutch physicist Dirk Coster and Hungarian chemist Georg von Hevesy. The find secured not only the periodic table's legacy but also the future of chemistry. Hafnium also came to represent a hard-won victory against those determined to undermine evidence-based discovery.

Dmitri Mendeleev's periodic table of elements, created in 1869, emerged from the realization that chemical elements such as oxygen and hydrogen share certain relationships. Mendeleev's contribution, and that of the German chemist Julius Lothar Meyer, working independently, provided an order for the elements, along with criteria for classifying them into neat groups. Remarkably, both Mendeleev's and Lothar Meyer's schemes were based on elements' subatomic structure — several decades before the discovery of electrons and protons.

When Mendeleev devised the periodic table's rough form, he started with 63 known elements. To make the table work, he had to leave gaps where as-yet undiscovered elements might be placed. These elements soon began to turn up. For example, the predicted 'element 68', gallium, was identified a few years later, in 1875. By 1914, just seven gaps remained.

A breakthrough occurred in 1913, when Henry Moseley, a British physicist, showed that elements could be arranged by their atomic number, or their number of protons. Moseley's work provided both a more accurate 'gap map' and a method for identifying elements from the spectra produced by exposing candidate elements to X-rays.



[The battle behind the periodic table's latest additions]

But the discovery (and naming) of element 72, hafnium, was anything but straightforward. The French chemist Georges Urbain originally proposed, in 1911, that element 72 belonged among the periodic table's rare earth elements, and named it celtium. But around a decade later, the Danish physicist Niels Bohr — who used quantum theory to develop a model of the atom in which electrons orbit the nucleus — predicted that element 72 would be among the transition metals, and closer to zirconium (element 40). This was finally confirmed by Coster and von Hevesy — both working at Bohr's lab in Copenhagen — who searched zirconium minerals for the element[2]. The duo named their discovery hafnium, after the Latin name for Copenhagen. They obtained the X-ray spectra in December 1922 and their paper[1] followed in January 1923.

But this was far from the end of the controversy, because Urbain stubbornly refused to give up, even though he had already had warning that the material he called celtium did not conform to the criteria for element 72. In 1914, Moseley and Urbain had collaborated on an unpublished X-ray study that failed to show that celtium was element 72. Urbain explained this away, saying that the X-ray method was simply not sensitive enough[3] — an assessment that the New Zealand-born physicist Ernest Rutherford, writing in *Nature*, agreed with[4]. Urbain also suggested that the Copenhagen team

were trying to take credit for his work[5]; in their response, Coster and von Hevesy refused to personalize the dispute and argued on the basis of their results[6].

Debate continued, with scientists from the Netherlands, Germany and Scandinavia on the side of the Copenhagen team, while those from France and the United Kingdom (who were backing a boycott of German science in the wake of the First World War) took the counter position. Hafnium was accepted by the International Union of Pure and Applied Chemistry only in 1930, a few years after the boycott formally ended. In France, celtium continued to occupy the place of element 72 among the rare earth elements until the early 1940s[2].

One hundred years on from hafnium's discovery, the periodic table remains both robust and relevant, even at a time when reams of data on an element can be accessed at the click of a mouse. The table offers, at a glance, a reference to how an element might behave in a chemical reaction, and clues to its similarity to other elements in a group.

So far, there are 118 confirmed elements, with the addition of four superheavy synthetic elements in 2015. But a time will surely come when there will be no new elements to make and the periodic table will reach its limits. As yet there is no consensus on when this is likely to happen or how big the last element will be, but chemists say its atomic number could exceed 170.

When that time comes, the periodic table will still remain, a map guiding scientists through the vastness of chemical space — all the molecules that have ever formed, and all those yet to be discovered, whether on Earth or elsewhere in the Universe. It's a tribute to the enduring values of international scientific cooperation and the steadfastness of researchers that an unremarkable transition metal, discovered 100 years ago in the aftermath of one of the world's greatest conflicts, made the periodic table what it is today.

Nature **613**, 413 (2023)

doi: https://doi.org/10.1038/d41586-023-00083-4

References

1. Coster, D. & Hevesy, G. *Nature* **111**, 79 (1923).
2. Kragh, H. *Centaurus* **23**, 275–301 (1980).
3. Frederick-Frost, K. M. *Chem. Int.* **41**(2), 23–27 (2019).
4. Rutherford, E. *Nature* **109**, 781 (1922).
5. Urbain, G. & Dauvillier, A. *Nature* **111**, 218 (1923).
6. Coster, D. & Hevesy, G. *Nature* **111**, 462–463 (1923).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00083-4

- EDITORIAL
- 18 January 2023

PhD training is no longer fit for purpose — it needs reform now

If researchers are to meet society's expectations, their training and mentoring must escape the nineteenth century.



Graduation day for PhD scholars at the University of Bologna, Italy. Doctoral assessment in most countries is largely unchanged since the nineteenth century.Credit: Roberto Serra/Iguana Press/Getty

These days, there's barely a world leader who doesn't talk up science. Prime Minister Narendra Modi was the star turn at the annual Indian Science Congress, held this month in Nagpur, where he exhorted the nation's

researchers to do the science needed to make India self-reliant. At last October's landmark Communist Party congress, Chinese Premier Xi Jinping set out his vision of how science and innovation could drive growth. And last August, US President Joe Biden signed the CHIPS and Science Act, which unlocks US$13.2 billion for semiconductor research and workforce development, in a bid to maintain the country's technological primacy.

In each case, the message to researchers is crystal clear: leaders see science as essential to national prosperity, well-being and, of course, competitiveness. So, is research fit for the challenge of advancing, refining or critiquing these goals? Not exactly. And it won't be until there is fundamental reform to the gateway to a research career: PhD training.

Training in trouble

As *Nature* and other publications have reported persistently, PhD training worldwide has been in trouble for some time. Surging inflation has eroded the already-meagre value of students' stipends in most countries, creating a cost-of-living crisis. Early-career researchers constantly report concerns about a chronic lack of support and poor-quality supervision, with senior researchers rarely trained in mentorship. Racism and discrimination are systemic in academic culture in many places.



Collection: Career resources for PhD students

Furthermore, PhD candidates are inadequately prepared for the cross-disciplinary working and large teams that characterize cutting-edge science today. This is especially true for careers outside academic research, where the overwhelming majority of PhD candidates will be heading.

It is not all bad. Universities in a small number of high-income countries have reformed, or are reforming, PhD assessment. But in most places, and especially in low- and middle-income countries, a candidate's work is still evaluated using a single-authored dissertation. This is 'defended' before a scholarly panel in what is still sometimes called a viva voce ('with living voice' in Latin), a nod to its nineteenth-century origins. And in many countries, candidates must publish in a journal before they get a PhD, something that critics say could fuel predatory publishing.

The system's strains have become more obvious because the number of people doing PhD training has been rising sharply. According to the 2022 book *Towards a Global Core Value System in Doctoral Education* — available as an open-access PDF; see go.nature.com/3zihyuk — the number of PhDs awarded in China more than doubled from 23,400 in 2004 to 55,011 in 2016 (and reached around 60,000 in 2019). India's numbers increased from 17,850 in 2004 to 25,095 in 2016; US figures climbed from 48,500 to 69,525 over the same period (see 'The rise of the PhD').

The rise of the PhD

The number of students awarded PhDs each year is expanding in a number of countries. In emerging economies, this follows growth in universal access to education.

Country	2004	2008	2012	2014	2016
Brazil	—	10,700	13,912	16,745	20,605

Country	2004	2008	2012	2014	2016
China	23,400	43,800	51,713	53,653	55,011
Germany	23,100	25,600	26,807	28,147	29,303
India	17,850	18,700	21,544*	21,830	25,095
Japan	16,900	17,300	15,911*	15,714	15,805
Mexico	—	—	—	—	6,970†
Russia	29,850	27,700	34,403	36,533	27,212
South Africa	—	—	—	—	3,057
South Korea	7,950	9,400	12,243	12,931	13,882
United Kingdom	15,300	16,600	20,438	25,020	27,366
United States	48,500	61,730	62,071	67,591	69,525

*2011; †2017. Source: M. Nerad *et al*. (eds) *Towards a Global Core Value System in Doctoral Education* (UCL Press, 2022). https://doi.org/10.14324/111.9781800080188

The doctorate updated

Too often, PhD training is still, at least conceptually, organized as it was after its development in and subsequent export from mid-nineteenth-century Germany. At that time, young scholars were attached to individual professors in a master–apprentice relationship, with the objective of safeguarding and advancing knowledge in individual disciplines.

That cannot continue if the next generation of scholars is to meet society's demands. There needs to be a revolution in the organization and funding of PhD training. And it must be on a similar scale to the shake-up that took place when school and university education stopped being a privilege enjoyed by relatively small numbers of people — those with wealth, or those training for occupations such as politics, religion or teaching. As education expanded to be open to everyone (at least in theory), it was no longer feasible to offer tuition in small groups where one person did most of the teaching, with little or no training. Countries introduced subject specialists as teachers; they limited class sizes, experimented with new educational technologies, collaborated on quality-assurance measures and drew on educational research to understand how students learn in different ways.

The editors and contributors of *Towards a Global Core Value System in Doctoral Education*, who are all experts in higher-education research, policy or practice, provide snapshots of the state of play in various countries, including all-important examples of innovation in PhD training. Some candidates are instructed in cohorts with more than one supervisor, so that students are less isolated and better protected if a relationship with a single supervisor goes bad. Some take additional courses of study, or have their research progress assessed periodically — the kind of guided approach that happens in education more broadly.

Reforming PhD training will be like turning a tanker. It will be slow. It will require planning and resources. But the end result must be a transformation on a massive scale, just as happened when primary, secondary and university education was reformed. If our leaders want scientists to work harder for society, they, in turn, need to work with research to ensure that PhD training finally exits the nineteenth century and joins the twenty-first.

Nature **613**, 414 (2023)

doi: https://doi.org/10.1038/d41586-023-00084-3

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00084-3

- WORLD VIEW
- 17 January 2023

Preprint review should form part of PhD programmes and postdoc training



Institutions owe it to young researchers to prepare them for careers outside academia. Preprint review is a perfect opportunity.

- Richard Sever

Few PhD students ultimately pursue the career for which they have been trained. A rapid expansion in PhD programmes over several decades without a concurrent increase in faculty numbers means that, in the United Kingdom, less than 5% of PhD students will become principal investigators (PIs), a Royal Society report found in 2010 (see go.nature.com/2r1y6pb). The numbers are not much better for the United States (R. C. Larson *et al.* *Syst. Res. Behav. Sci.* **31**, 745–750; 2014).

The rest go into industry, pursue science writing, editing, patent law or other science-related careers, or leave science entirely. They generally do so armed with just a thesis and one or two research papers on a rather arcane

topic, but little other tangible evidence of their abilities. We need a recognized, equitable way for PhD graduates to demonstrate the transferable skills they have gained.

For me, that way is to train them in preprint review.

I was once one of those postgraduates, choosing to leave the bench with little on my CV other than grades and degrees. In the years since, both in my current role as assistant director of Cold Spring Harbor Laboratory Press and in previous positions, I have often had to sift through dozens of CVs when hiring people for editorial positions.



[‘I don’t want this kind of life’: graduate students question career options]

Occasionally, interviewers like me see relevant experience on a CV, when someone has worked on a student journal, for example, or written a newspaper article, that helps us to decide whether to shortlist them. But internships outside academia and other writing opportunities are generally the preserve of a fortunate few.

Peer reviewing preprints would guarantee young researchers some concrete outputs that illustrate their ability to critique work, write about science and discuss subjects outside their immediate focus of research. By building such training into our scientific institutions, rather than relying on outside

opportunities to which many do not have access, we can create a fairer system in which not just the well-connected can demonstrate their abilities.

Peer review is an opportunity to demonstrate transferable skills, but most of the institutions that grant PhDs don't train people in peer review. Journal peer review tends to be confined to PIs; at most, a student might work with their PI to learn the ropes. Even then, they are unlikely to be credited. The process takes place mainly behind closed doors and reports are not publicly registered, so job applicants cannot cite them as evidence of their abilities.

The rise in preprinting provides a fantastic opportunity for PhD students and postdoctoral researchers to engage in peer review, make the reviews public and use them as career currency.



To fix peer review, break it into stages

Since we launched bioRxiv in 2013, increasing numbers of biologists have been posting preprints online, following the lead of physicists with the arXiv server, founded in 1991. These have yet to be scrutinized for errors, flawed logic or extra work that is needed, so they are ideal subjects for young investigators who want to demonstrate their criticism skills. Many student journal clubs increasingly discuss preprints because the work has not yet been formally evaluated and revised. Importantly, there is no gatekeeping: anyone who wants to peer review a preprint can.

Institutions could build their own platforms for students' preprint reviews, modelled on PREreview and preLights. PREreview is an open-source platform that enables users to compose, read and endorse long-form peer reviews of preprints. PreLights is a preprint-highlighting service in which early-career researchers select, summarize and comment on preprints. Both assign digital object identifiers (DOIs) to preprint reviews, and the reviews can be automatically linked to a scientist's ORCID (Open Researcher and Contributor ID) account. Preprint reviews thus have the potential to serve not only as a career currency, but as one that can be authenticated.

For institutions motivated to start peer-review training, PREreview spells out how to peer review constructively (see go.nature.com/3gzsex0) and ASAPbio's working group on preprint feedback has collected useful guides (see go.nature.com/3qdjapt). Individual labs and departments can also begin teaching peer review, as some already do. Ultimately, however, institutions should initiate this rather than depending on piecemeal efforts by individual PIs.

A few institutions have begun to offer formal training in peer review. New York University and the University of California, San Francisco, both run courses, for example.

Some might argue that preprint review is an unnecessary distraction — another hoop through which we force already busy people to jump — that's not necessarily of value to all trainees. But in my view, these new outputs would provide multidimensional information about candidates' qualities for potential employers in and outside academia. For researchers who stay in academia, training will improve their peer reviews. It might even prevent them from becoming the dreaded 'Reviewer 3', the incompetent critic who makes the process unbearable for authors.

PhD students and postdocs are often referred to as trainees, but what career are we training them for? In a 2022 *Nature* survey, only 32% of graduate students said that their supervisors had useful advice for careers outside academia (*Nature* **611**, 413–416; 2022). We need to do better. Ensuring that students have material evidence of transferable skills would be an important step forwards.

Nature **613**, 415 (2023)

doi: https://doi.org/10.1038/d41586-023-00085-2

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00085-2

| Section menu | Main menu |

News in Focus

- **Trial settles debate over best design for mRNA in COVID vaccines** [12 Jan 2023]
  • The biggest defender of 'unmodified' mRNA for vaccines against infectious disease has a change of heart.
- **Will Brazil's President Lula keep his climate promises?** [06 Jan 2023]
  • The new president has already taken steps to protect the Amazon, but researchers call for more action.
- **Historic US research strike ends — but energizes a movement** [11 Jan 2023]
  • As University of California researchers return to work, union organizers say this is just the beginning of a nationwide push for better working conditions.
- **Abstracts written by ChatGPT fool scientists** [12 Jan 2023]
  • Researchers cannot always differentiate between AI-generated and original abstracts.
- **China's COVID wave has probably peaked, model suggests** [16 Jan 2023]
  • But a lack of data is obscuring the true impact of the outbreak.
- **These 'green pea' galaxies might have helped to end the Universe's dark age** [13 Jan 2023]
  • JWST has spotted three tiny, distant galaxies that could solve a long-standing cosmic mystery.
- **2022 was a record year for space launches** [11 Jan 2023]
  • 180 rockets lifted off successfully, with SpaceX driving the pace.
- **How your first brush with COVID warps your immunity** [18 Jan 2023]
  • The immune system responds more strongly to the strain of a virus that it first met, weakening response to other strains. Can this 'imprinting' be overcome?

- NEWS
- 12 January 2023

Trial settles debate over best design for mRNA in COVID vaccines

The biggest defender of ‘unmodified’ mRNA for vaccines against infectious disease has a change of heart.

- Elie Dolgin



Ingredients for the mRNA vaccine made by CureVac.Credit: Guillaume Souvant/AFP via Getty

It’s only a small chemical tweak, but it has triggered huge division in the mRNA research community.

Proponents argue that a simple chemical modification to the backbone of mRNA is crucial to the success of mRNA vaccines, allowing them to trigger a potent immune response without massive side effects. Exhibit A: the blockbuster COVID-19 vaccines made by Pfizer–BioNTech and Moderna, both of which have this modification. But some mRNA manufacturers have long held that unmodified mRNA elicits a superior immune response to pathogens — a rift with implications for mRNA vaccine development.

Now the biggest champion of the unmodified approach has changed its tune. CureVac — which is one of the world’s oldest mRNA specialty companies and is based in Tübingen, Germany — is ditching unmodified mRNA, and embracing the modified version for its entire infectious-disease vaccine portfolio after disappointing results from COVID-19 vaccine trials.

Modified mRNA is the “best-performing technology for prophylactic vaccines”, Franz-Werner Haas, CureVac’s outgoing chief executive, said in a conference call last week to announce the company’s strategic shift — a retreat from a position it has held since its founding in 2000.

The company’s about-face should help put to rest the decades-long debate among scientists over which type of mRNA is best to protect against influenza, SARS-CoV-2 and other viral threats. CureVac’s pivot in the respiratory-disease arena leaves only a few late-stage products reliant on unmodified mRNA, all of which are COVID-19 vaccine candidates under development in Asia.

The mod squad

Vaccines designed around mRNA work by giving the body instructions on how to make viral proteins. Those proteins then train the immune system to produce antibodies that can disarm invading pathogens. But the process only works if the immune system doesn’t see the vaccine mRNA itself as a threat.

That’s where modifications come in. In the mid-2000s, scientists discovered that swapping one of RNA’s four basic building blocks, called nucleotides,

for a related analogue helped to make the mRNA less conspicuous to the immune system. Vaccines could then be given at higher doses, allowing for maximal antibody responses with fewer of the icky side effects — fevers, chills, aches — that result from immune reactions to the mRNA itself.



Fast-evolving COVID variants complicate vaccine updates

For decades, CureVac thought that it could achieve the same effect by fine-tuning the mRNA sequence to avoid tripping the immune alarm. The company claimed that its 'optimized' mRNA could elicit the full breadth of desired immune responses without any major safety issues[1]. But in late-stage testing, CureVac's experimental jab didn't yield enough antibodies to thwart emerging variants of SARS-CoV-2[2]. The general consensus: CureVac's dose was too low but couldn't be raised because of tolerability concerns.

The company created a next-generation jab that, although still built around unmodified mRNA, seemed to generate more robust immune responses at the same dose used for its earlier vaccine[3]. The product entered clinical trials in March 2022. But around the same time, a shift in mindset was taking hold at the company.

As Haas told *Nature* in an interview that month, "There should not be a kind of religious way to see" the issue of mRNA modifications. CureVac and its

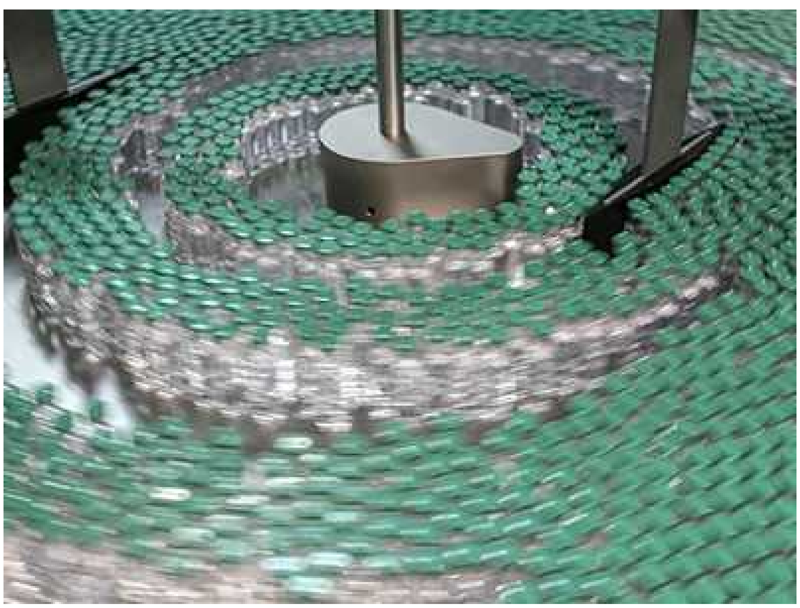
co-development partner, GSK of London, decided to give modified mRNA a real try, with the idea being, as Haas put it, to “let the data talk”.

In August, another COVID-19 vaccine candidate from CureVac entered clinical trials, this one updated to reflect the circulating Omicron variant of SARS-CoV-2 — and now with modified mRNA.

Chemical conversion

On 6 January, CureVac reported preliminary-trial data comparing the two next-gen jabs. At the same dose, a shot of the modified mRNA elicited about the same number of coronavirus-foiling antibodies as the next-generation vaccine based on unmodified mRNA. But importantly, it also caused many fewer side effects, allowing CureVac to safely up the dosage for maximal immune protection.

The data in favour of modifications were “very compelling”, says Phil Dormitzer, global head of vaccines research and development at GSK. “Modified RNA has really proven itself.” CureVac, in collaboration with GSK, is now moving forward with high-dose modified mRNA for both of its COVID-19 and flu vaccine programmes.



Could computer models be the key to better COVID vaccines?

CureVac is not the first firm to run trials of mRNA vaccine candidates, with and without modifications. Other companies, including BioNTech in Mainz, Germany, and Sanofi, headquartered in Paris, have also done so — but none had disclosed their full results. “That’s why the CureVac study is so helpful,” says vaccinologist Norbert Pardi at the University of Pennsylvania Perelman School of Medicine in Philadelphia.

The data might be preliminary, but “this is the first head-to-head comparison led by the same company”, he says. And, according to Pardi, it clearly shows that “modified RNA is the way to go”.

Another long-time proponent of ‘natural’ mRNA, the company Translate Bio, now a subsidiary of Sanofi, ditched the approach last year. That leaves only a handful of newer mRNA players still advocating for unmodified mRNA in respiratory vaccines. Among them is the company Abogen Biosciences of Suzhou, China, which is furthest along with its late-stage trials of a COVID-19 vaccine developed using unmodified mRNA. Few scientists expect the vaccine to perform much better than CureVac’s first product did.

“The industry has pretty much all aligned with using modified nucleotides,” says Michael Heartlein, a co-founder of Translate Bio.

A role for natural products

That’s true for infectious diseases at least, and with current formulations. There are other, as-yet unproven ways to deliver mRNA vaccines that might not need the chemical switcheroo. And in the realm of oncology, unmodified mRNA-based vaccines might be useful for ‘therapeutic vaccines’, which treat disease, rather than prevent it.

Cancer vaccines help the immune system to recognize tumour cells so that killer T cells can mount an attack. In this context, natural-mRNA vaccines’ tendency to trigger the production of immune-revving molecules can be beneficial.

The same molecules could suppress the types of antibody response needed to combat infectious-disease agents, notes Katalin Karikó, the biochemist

who first demonstrated the importance of mRNA modifications for avoiding unwanted immune responses, and who now works a consultant for BioNTech. However, she adds that for cancer, killer immune cells are the key.

A test should come soon: later this year, both BioNTech, whose cancer vaccines use unmodified mRNA, and Moderna, which uses modified, will report full data on their vaccines' performance in treating melanoma.

Nature **613**, 419-420 (2023)

doi: https://doi.org/10.1038/d41586-023-00042-z

References

1. Lutz, J. *et al. NPJ Vaccines* **2**, 29 (2017).

2. Kremsner, P. G. *et al. Lancet Infect. Dis.* **22**, 329–340 (2022).

3. Gebre, M.S. *et al. Nature* **601**, 410–414 (2022).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00042-z

- NEWS
- 06 January 2023

Will Brazil's President Lula keep his climate promises?

The new president has already taken steps to protect the Amazon, but researchers call for more action.

- Meghie Rodrigues



President Luiz Inácio Lula da Silva took office on 1 January in Brazil.Credit: Evaristo Sa/AFP via Getty

Expectations were high this week as Brazil's new president, Luiz Inácio Lula da Silva, took office. Lula campaigned on promises to help the world fight climate change by protecting the Amazon rainforest, which sequesters a large portion of global carbon emissions. Under his predecessor, Jair Bolsonaro, deforestation in the Brazilian Amazon reached its highest level since 2008, threatening efforts to reduce worldwide emissions and preserve one of Earth's richest biodiversity hotspots.



Brazil election: Scientists cheer Lula victory over Bolsonaro

Global leaders and scientists are waiting to see whether Lula will be able to fulfil his pledges. In November 2022, at the COP27 international climate summit in Sharm El-Sheik, Egypt, he was greeted with cheers as he proclaimed that "Brazil is back" and committed to reaching net-zero deforestation by 2030.

He will be fighting an uphill battle, say analysts and researchers who spoke to *Nature*. Several of Bolsonaro's allies were elected to the Brazilian Congress last year, so Lula will need to forge his own alliances to pass climate legislation. He will also have to contend with Bolsonaro's legacy, finding ways to reverse actions the former leader took against environmental protection.

“Bolsonaro weakened monitoring systems for protected areas, reduced budgets for environmental institutions” and replaced researchers at those organizations with military staff who had no scientific expertise, says Marina Silva, Brazil’s new environment minister. She had the same role from 2003 to 2008, during Lula’s first and second terms as president. Under the Bolsonaro administration, there was an “overall policy blackout”, says Silva, who was also part of an advisory group that helped Lula to transition into office after winning the presidency in late October 2022.

A starting place

Lula has already signed several climate- and environment-related decrees in his first week on the job. One re-establishes the Amazon Fund, an international mechanism frozen by the Bolsonaro administration that finances efforts to reduce deforestation. Lula also revoked a 2022 edict, signed by Bolsonaro, that sought to expand and legalize small-scale ‘wildcat’ gold mining, which strips land of vegetation, pollutes waterways and is often carried out illicitly in Indigenous territories in the Amazon. Marina Silva got to work quickly, too, announcing the creation of a National Authority for Climate Security, an office that — if approved by the Brazilian Congress — will oversee the execution and implementation of the country’s climate policy.



Illegal gold mining has stripped vegetation from the Amazon.Credit: Mauro Pimentel/AFP via Getty

According to Natalie Unterstell, president of the Talanoa Institute, a think tank focused on climate policy in Rio de Janeiro, Brazil, even more action is needed to restore Brazil's climate leadership. The Lula administration should start by restructuring the Ministry of the Environment and Climate Change, she says, and by restoring the mandate of the agencies it houses — all of which have been destabilized over the past four years. Under Bolsonaro, the budget for Brazil's environment agencies dropped by more than 70%, compared with 2014 (when the budget was at its peak), leaving it at its lowest level in 17 years.

Another crucial step the new administration should take, Unterstell adds, is to "put law enforcement back in place" to protect against environmental crime. In 2021, the country issued the lowest number of fines for logging in two decades, according to the Brazilian newspaper *O Estado de São Paulo*, even though the Amazon experienced its highest rates of deforestation in more than a decade.



[Bolsonaro's troubled legacy for science, health and the environment]

The rainforest takes in about one-quarter of all the carbon dioxide absorbed by land on Earth. Research shows[1], however, that deforestation is transforming the region from a carbon sink into an emissions source, and scientists worry that in about a decade, the Amazon could reach a tipping point — when disrupted rain cycles and biodiversity loss could convert it into a savannah-like area — if logging goes unchecked.

To put Brazil on a path to net-zero deforestation, the new government should start by strengthening surveillance of the rainforest's health and use the data to influence environmental policy, Unterstell says. Restoring funding and staff to Brazil's National Institute for Space Research (INPE) — which uses satellites to monitor tree cover and greenhouse-gas emissions, and was gutted by Bolsonaro — would be an important step in that direction, she argues.

Great expectations

If Brazil's new government is to reconstruct the country's environmental policy, science must play a crucial part, says former INPE director Ricardo Galvão, who, like Silva, was part of the transition team advising Lula. Galvão says that the transition team recommended the creation of an office

devoted to the Amazon, to be placed in the Ministry of Science, Technology and Innovation. Lula followed through on this advice, establishing the Sub-secretariat of Science and Technology for the Amazon during his first day in office, on 1 January. The idea is for it to "help oversee scientific development in the region and put forth biodiversity-based sustainable development", Galvão adds. This means finding ways to allow economic development to occur in the Amazon without wiping out plant and animal species, by harvesting non-timber forest products such as plant-based oils or fruits, for example.



What a new president in Brazil could mean for science

All of Lula's climate plans won't be possible without funding, however, and that's what scientists will be watching for, says Paulo Artaxo, a physicist at the University of São Paulo. An increase in science and technology funding, as well as a rise in research grants — hopefully within Lula's first 100 days in office — will be some of the markers of success, Artaxo adds.

Cuts to science budgets have been felt widely throughout Brazil — and beyond. Erika Berenguer, an ecosystems researcher at the University of Oxford, UK, is part of the Sustainable Amazon Network, which collects scientific evidence as a way to strengthen sustainability in the Amazon. The network, made up of institutions inside and outside Brazil, has had

difficulties with staffing and training because of the cuts. “Both benefits and lack of opportunity are shared among us all,” Berenguer says.

Silva is optimistic that the Lula administration will turn things around. As challenging as the current situation is, she says, the new government has a roadmap it can follow: during her stint in the 2000s as environment minister, she created the Action Plan for the Prevention and Control of Deforestation in the Legal Amazon, which helped to slash deforestation by 83% in the region between 2004 and 2012. The challenges are different this time, but “Brazil knows what to do and how to do it”, Silva says. “We’ll build on that success.”

Nature **613**, 420-421 (2023)

doi: https://doi.org/10.1038/d41586-023-00011-6

References

1. Gatti, L. V. *et al. Nature* **595**, 388–393 (2021).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00011-6

- NEWS
- 11 January 2023
- Correction 12 January 2023

Historic US research strike ends — but energizes a movement

As University of California researchers return to work, union organizers say this is just the beginning of a nationwide push for better working conditions.

- Jeff Tollefson



The University of California strike began on 14 November last year and ended six weeks later.Credit: Mel Melcon/*Los Angeles Times* via Getty

Research has resumed across the University of California (UC) system, following a breakthrough in negotiations late last month that brought [the largest higher-education strike in US history] to a close. Tens of thousands of graduate students, postdoctoral researchers and other academic staff are heading back to work with higher wages and more benefits because of the strike, but that's not all: the revolt has injected fresh momentum into a growing unionization movement at university campuses across the United States.



[Thousands of academics strike in California: how is research affected?]

"This shows that massive strikes in higher education are possible, and that people can win significant improvements in their working conditions," says Rebecca Givan, co-director of the Center for Work and Health at Rutgers University in New Brunswick, New Jersey, and chair of the union representing academic employees there. "Academic workers everywhere are taking note."

Although the deals to end the strike left some UC workers disappointed, advocates and scholars say that the six-week stand-off represents a landmark achievement for a growing labour movement. Some 48,000 employees across all 10 of the university's campuses and the Lawrence Berkeley National Laboratory will receive expanded benefits and wage increases ranging from 20% to 80%, which supporters say will help offset rising costs for housing and health care. Union organizers also proclaimed it the very

first “research strike”, as graduate students, postdoctoral researchers and non-tenure-track academics united to bring science at their institutions to a standstill.

There was a perception that “the work we do in the labs is not labour”, says Rafael Jaime, a scholar in English literature at the University of California, Los Angeles (UCLA), and president of the union representing graduate students and undergraduate workers. “But we have changed that perception.”

The word is spreading: union representatives involved in strikes at UC and other universities say that they are fielding queries from researchers at various academic institutions about how to organize unions, conduct negotiations and implement strikes. Some are even preparing workshops in the coming months to accommodate demand.

“This is a movement about hope,” says Johannah King-Slutzky, a graduate student in literature at Columbia University in New York and a union organizer there, who is helping with the workshops. “When you see a success on one campus, you feel energized to take similar steps on your own.”

Organizing labour

Although unusual in its size and scope, the UC strike is just the latest example of a growing protest movement in academia. Over the past year, strikes have been held at several US universities, including Indiana University in Bloomington, and Columbia University. Similar demands for higher wages and better working conditions have accompanied strikes in places such as Canada and the United Kingdom, although these countries have a longer and more ingrained history of organized labour.



PhD students face cash crisis with wages that don’t cover living costs

Union organizers say that the effects of such strikes are already bringing changes at some US institutions. Both Princeton University in New Jersey, and the University of Pennsylvania in Philadelphia offered historic — and unprompted — increases in research stipends for graduate students last year, citing the need to attract and maintain doctoral students. “It’s hard to make a causal link, but the breadth of evidence is suggestive that the strikes are having major repercussions across academia in the United States,” says Connor Jackson, a graduate researcher in agricultural economics at the Lawrence Berkeley laboratory.

As it stands, academia is now home to some of the largest growth in unionization across all sectors of the US economy. Nearly 150 unions, representing more than 600,000 academic workers, have signed a statement developed by the organization Higher Education Labor United that advocates wall-to-wall union protection for higher-education workers across the country. And unionization efforts are expanding at universities across the country, union officials say.

The surge in the academic labour movement has been decades in the making, with reductions in public funding pushing universities to rely on higher tuition fees, cheap labour and part-time or adjunct faculty members to continue functioning. Rising costs for housing and health care have only compounded the problem — particularly in places such as California, where

the cost of living has skyrocketed in recent years. At the same time, the period that graduate students and postdoctoral workers spend in economic hardship has lengthened, as tenure-track professorships have dwindled.

“It’s just getting harder and harder to justify poverty wages by saying that grad workers are apprentices, with some kind of golden ticket for a tenure-track position,” says Steve Striffler, director of the Labor Resource Center at the University of Massachusetts, Boston.

Missed opportunity?

Academics say this is just the beginning. The rising expectations of an emboldened labour movement were on full display on 23 December, when more than 35% of the members in two unions representing UC graduate students voted against accepting university officials’ offer and ending the strike.



Massive strikes at UK universities over ‘unsustainable’ working conditions

One of the organizers of the vote-no campaign was Dylan Kupsh, a graduate researcher in computer science at UCLA. Kupsh was in close contact with union organizers at Columbia University, where student workers rejected an

initial contract proposal and went on to secure further concessions after a ten-week strike that ended last January.

In the end, UC graduate students received a range of new benefits, including increased childcare subsidies; protections against bullying, discrimination and harassment; and a new schedule for salaries. Incoming graduate students, for example, will see their annual salary increase from around US$22,000 to $30,500. “We could have won a lot more, and it’s sad that we didn’t get there,” Kupsh says. “We’re going to have to repeat in another 2.5 years.”

For Barry Eidlin, a sociologist at McGill University in Montreal, Canada, who studies the labour movement, the scale of the vote-no campaign is yet another sign of changing expectations in academia. “In the past, academic workers have felt like they should just keep their heads down and be grateful they have a job,” he says. “The idea that people now expect more, and are willing to fight for more, seems to me a welcome shift in perspective.”

Nature **613**, 422-423 (2023)

doi: https://doi.org/10.1038/d41586-023-00049-6

Updates & Corrections

- **Correction 12 January 2023**: An earlier version of this story gave the wrong location for Princeton University. The text has been corrected.

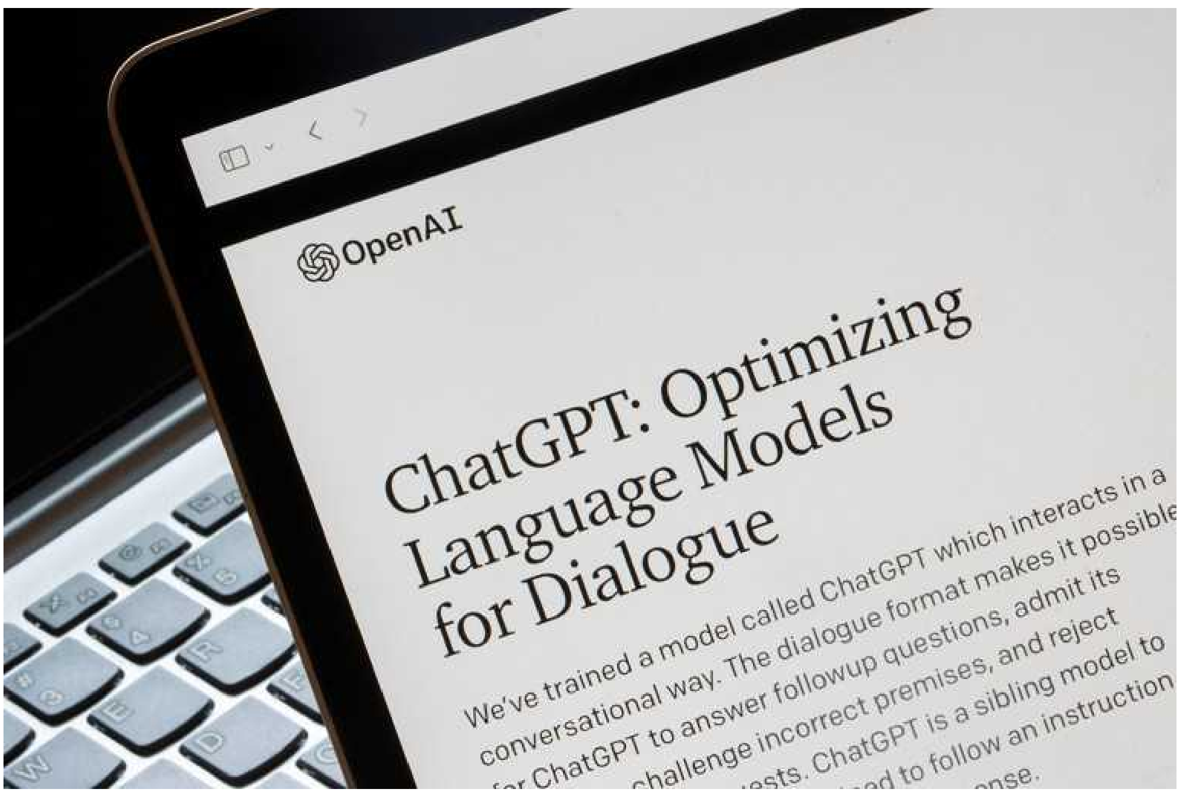
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00049-6

- NEWS
- 12 January 2023

Abstracts written by ChatGPT fool scientists

Researchers cannot always differentiate between AI-generated and original abstracts.

- Holly Else



Scientists and publishing specialists are concerned that the increasing sophistication of chatbots could undermine research integrity and accuracy.Credit: Ted Hsu/Alamy

An artificial-intelligence (AI) chatbot can write such convincing fake research-paper abstracts that scientists are often unable to spot them, according to a preprint posted on the bioRxiv server in late December[1]. Researchers are divided over the implications for science.

“I am very worried,” says Sandra Wachter, who studies technology and regulation at the University of Oxford, UK, and was not involved in the research. “If we’re now in a situation where the experts are not able to determine what’s true or not, we lose the middleman that we desperately need to guide us through complicated topics,” she adds.

The chatbot, ChatGPT, creates realistic and intelligent-sounding text in response to user prompts. It is a ‘large language model’, a system based on neural networks that learn to perform a task by digesting huge amounts of existing human-generated text. Software company OpenAI, based in San Francisco, California, released the tool on 30 November, and it is free to use.

Since its release, researchers have been grappling with the ethical issues surrounding its use, because much of its output can be difficult to distinguish from human-written text. Scientists have published a preprint[2] and an editorial[3] written by ChatGPT. Now, a group led by Catherine Gao at Northwestern University in Chicago, Illinois, has used ChatGPT to generate artificial research-paper abstracts to test whether scientists can spot them.

The researchers asked the chatbot to write 50 medical-research abstracts based on a selection published in *JAMA*, *The New England Journal of Medicine*, *The BMJ*, *The Lancet* and *Nature Medicine*. They then compared these with the original abstracts by running them through a plagiarism detector and an AI-output detector, and they asked a group of medical researchers to spot the fabricated abstracts.

Under the radar

The ChatGPT-generated abstracts sailed through the plagiarism checker: the median originality score was 100%, which indicates that no plagiarism was detected. The AI-output detector spotted 66% the generated abstracts. But the human reviewers didn't do much better: they correctly identified only

68% of the generated abstracts and 86% of the genuine abstracts. They incorrectly identified 32% of the generated abstracts as being real and 14% of the genuine abstracts as being generated.

“ChatGPT writes believable scientific abstracts,” say Gao and colleagues in the preprint. “The boundaries of ethical and acceptable use of large language models to help scientific writing remain to be determined.”

Wachter says that, if scientists can’t determine whether research is true, there could be “dire consequences”. As well as being problematic for researchers, who could be pulled down flawed routes of investigation, because the research they are reading has been fabricated, there are “implications for society at large because scientific research plays such a huge role in our society”. For example, it could mean that research-informed policy decisions are incorrect, she adds.

But Arvind Narayanan, a computer scientist at Princeton University in New Jersey, says: “It is unlikely that any serious scientist will use ChatGPT to generate abstracts.” He adds that whether generated abstracts can be detected is “irrelevant”. “The question is whether the tool can generate an abstract that is accurate and compelling. It can’t, and so the upside of using ChatGPT is minuscule, and the downside is significant,” he says.

Irene Solaiman, who researches the social impact of AI at Hugging Face, an AI company with headquarters in New York and Paris, has fears about any reliance on large language models for scientific thinking. “These models are trained on past information and social and scientific progress can often come from thinking, or being open to thinking, differently from the past,” she adds.

The authors suggest that those evaluating scientific communications, such as research papers and conference proceedings, should put policies in place to stamp out the use of AI-generated texts. If institutions choose to allow use of the technology in certain cases, they should establish clear rules around disclosure. Earlier this month, the Fortieth International Conference on Machine Learning, a large AI conference that will be held in Honolulu, Hawaii, in July, announced that it has banned papers written by ChatGPT and other AI language tools.

Solaiman adds that in fields where fake information can endanger people's safety, such as medicine, journals may have to take a more rigorous approach to verifying information as accurate.

Narayanan says that the solutions to these issues should not focus on the chatbot itself, "but rather the perverse incentives that lead to this behaviour, such as universities conducting hiring and promotion reviews by counting papers with no regard to their quality or impact".

Nature **613**, 423 (2023)

doi: https://doi.org/10.1038/d41586-023-00056-7

References

1. Gao, C. A. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.12.23.521610 (2022).
2. Blanco-Gonzalez, A. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2212.08104 (2022).
3. O'Connor, S. & ChatGPT *Nurse Educ. Pract.* **66**, 103537 (2023).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00056-7

- NEWS
- 16 January 2023
- Correction 16 January 2023

China’s COVID wave has probably peaked, model suggests

But a lack of data is obscuring the true impact of the outbreak.

- Dyani Lewis



Hospitals in China have been overwhelmed by people with COVID-19.Credit: Hector Retamal/AFP via Getty

China’s massive COVID-19 outbreak probably peaked in late December, according to a preliminary analysis on the number of infections late last year and data on travel between cities. But public-health experts are frustrated by a lack of official data on the magnitude and severity of the outbreak.

For close to three years, China’s strict zero-COVID policy kept infections at bay. But since President Xi Jinping abruptly changed course and abandoned the policy in December, the Omicron variant of the coronavirus has spread largely unmitigated across the country.

Yet there are signs that the current wave of infections has already crested in many parts of China, according to Shengjie Lai, an infectious-diseases modeller at the University of Southampton, UK. In late December, Lai simulated the number of infections in different regions of China by combining information on how the variant was spreading in October and November 2022 with data on travel between cities across the country.

According to Lai’s analysis, shown to *Nature* but not yet published or peer reviewed, close to half of China’s cities experienced a peak in infections between 10 December and 31 December. For a further 45% of cities, the peak is predicted to occur in the first half of January.

Online search

This aligns with peak times that Lai estimated using searches for terms such as ‘fever’ and ‘COVID’ on the Internet search platform Baidu. It also fits with reports on the extent of infections in certain cities and provinces. For instance, on 21 December, the deputy director of the Chinese Center for Disease Control and Prevention (China CDC) in Beijing said during a briefing that more than 250 million people — some 18% of the population — had already been infected. And in large cities such as Beijing and Sichuan, more than 50% of residents had been infected, he said.

Meanwhile, in Henan — China’s third most populous province — an official from the province’s health commission said at a press conference that nearly 90% of Henan’s population had been infected by 6 January. The estimate was probably determined from online surveys that local health authorities

are conducting around the country, according to Lai and others, because more people are being infected than are being tested.

Christopher Murray, the head of the Institute for Health Metrics and Evaluation in Seattle, Washington, is sceptical of such infection estimates, because there has been no transparency about how they were made. Modelling carried out by the institute, and published on 16 December, suggested that the country's outbreak might not peak until as late as April.

But epidemiologist Jodie McVernon at the Doherty Institute in Melbourne, Australia, says that it makes more sense that the outbreak has already peaked, given how rapidly the Omicron variant spreads. "The idea that it would still be growing over the next few months just doesn't make any sense," she says.

Rural China

Because the virus has already spread rapidly across China, fears that city-dwellers could unleash outbreaks in rural parts of the country during *chun yun* — the 40-day Lunar New Year travel period that commenced on 7 January — are probably overblown, says Lai. "The virus has already spread to the rural areas," he says, and his modelling suggests that curbing travel would do little to alter the outbreak.

People in rural China could still be hit hard by severe illness and deaths, says Xi Chen, a health economist at Yale University in New Haven, Connecticut. Roughly 40% of China's older population live in rural regions that lack access to larger hospitals that are better equipped to treat severe cases, he says.

Hui Jin, a public-health researcher at Southeast University in Nanjing, China, says that vaccine hesitancy is common among older people, but that vaccination rates on the Chinese mainland have increased over the past year. According to data released by the State Council's joint prevention and control mechanism, 86% of those over the age of 60, and 66% of those older than 80, had been fully vaccinated by late November.

Death toll unknown

Epidemiologists are keen to know how many people are dying from COVID-19 in China. On 14 January, China's National Health Commission reported that close to 60,000 people have died from COVID-19 since 8 December 2022. The figure includes 5,503 deaths from COVID-19-related respiratory failure, as well as more than 54,000 deaths in people with COVID-19 and other underlying conditions. But the number covers only people who died in hospital.

On 29 December, Zunyou Wu, chief epidemiologist at the China CDC, told a press conference that the agency was already working to assess excess mortality — a measure of how many more deaths occurred than would otherwise be expected — and has plans to publish the data.

But Murray worries that the true toll of the COVID-19 outbreak in China will be difficult to ascertain because the country's system for recording births and deaths is incomplete.

Ariel Karlinsky, an economist at the Hebrew University in Jerusalem and a member of the World Health Organization's technical advisory group on COVID-19 mortality assessment, says that excess-mortality data could indicate the outbreak's death toll. But it could be another year before such data can be calculated, because deaths in 2023 will probably not be known until they are reported in China's annual statistical yearbook in January 2024.

Although the current infection wave might be ending, McVernon notes that China is likely to face a cycle of these surges over the next year. "In having these really big, massive waves, there's a lot of exposures all at once, and then there's a lot of waning all at once," she says.

Nature **613**, 424-425 (2023)

doi: https://doi.org/10.1038/d41586-023-00075-4

Updates & Corrections

- **Correction 16 January 2023**: A previous version of this story misspelt the names of Xi Jinping and Jodie McVernon.

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00075-4

- NEWS
- 13 January 2023

These ‘green pea’ galaxies might have helped to end the Universe’s dark age

JWST has spotted three tiny, distant galaxies that could solve a long-standing cosmic mystery.

- Shannon Hall



JWST has helped astronomers to discover distant galaxies (right) that are similar to ‘green pea’ galaxies (left, imaged by the Sloan Digital Sky Survey) from the nearby Universe.Credit: SDSS and NASA, ESA, CSA, and STScI

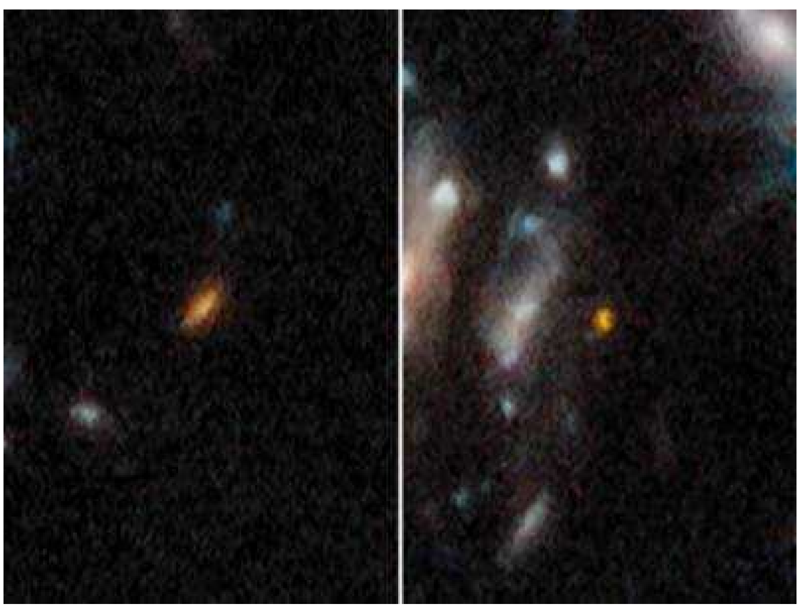
Astronomers using the James Webb Space Telescope (JWST) have found three tiny galaxies that might have helped to trigger one of the greatest

remodels in cosmic history.

The findings, presented at a meeting of the American Astronomical Society (AAS) in Seattle, Washington, and published last week[1], could explain reionization — the period when harsh radiation tore apart a 'fog' of hydrogen atoms hanging over the Universe, rendering stars and galaxies visible for the first time.

A cosmic blackout

In the immediate aftermath of the Big Bang, the Universe was blindingly bright. The residual heat was so great that electrons could not join with protons to form atoms. Instead, the Universe was a seething plasma — a dense glowing gas of electrically charged (or ionized) particles that scattered light, much like a fluorescent light bulb.



JWST spots some of the most distant galaxies ever seen

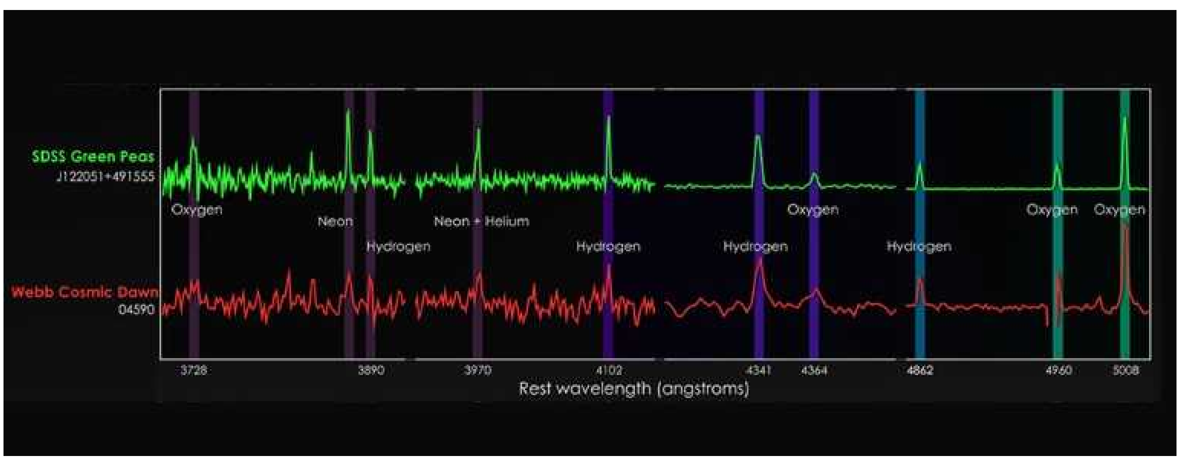
After about 380,000 years, the expanding Universe had cooled enough for hydrogen atoms to form. Some of those atoms eventually coalesced into the first stars and galaxies. But the surrounding blanket of hydrogen absorbed their light — similar to how a thick fog can block a vehicle's headlights.

Eventually, this cosmic dark age ended. Energetic radiation broke the intergalactic hydrogen atoms apart — turning them back into lone protons and electrons — in a process called reionization. But ionizing all of the matter between galaxies would take a vast amount of energy, and astronomers have long argued over the exact driver. Perhaps it was starlight from the earliest galaxies. Or maybe the event was caused by supermassive black holes — which tug matter towards them and heat it up.

“One of the big cosmological questions is, what is the cause of reionization?” says Trinh Thuan, an astronomer at the University of Virginia in Charlottesville.

Galactic peas

On 11 July 2022 — about six months after JWST launched — the observatory delivered the deepest and sharpest image of the distant Universe yet seen. The image, which centres on a galaxy cluster called SMACS 0723, contains thousands of galaxies fainter than any seen before. JWST could capture it because it analyses infrared light; as the Universe expands, its most distant galaxies appear to shift out of the visible part of the electromagnetic spectrum and into the infrared. Of the thousands of galaxies in the SMACS 0723 image, researchers decided to make follow-up spectroscopic observations of three that looked as though they might be especially far away.



The emissions spectra of green pea galaxies are similar to those of galaxies from the early Universe, spotted by JWST.Credit: NASA’s Goddard Space

Flight Center/Rhoads et al. 2023

But when James Rhoads, an astronomer at the NASA Goddard Space Flight Center in Greenbelt, Maryland, and his colleagues first saw the trio's spectra, they realized that the galaxies looked like objects normally found nearby. They had an uncanny resemblance to 'green pea' galaxies, oddities first spotted by a citizen-science project called Galaxy Zoo and reported[2] in 2009.

Green pea galaxies were named because of their colour and small size. They are just 5% of the size of the Milky Way, and 1% of the mass. But don't let their small size fool you: they churn out stars at an enormous rate — roughly 100 times as fast as astronomers would expect, given their mass. They also seem to contain relatively few heavy elements. Their greenish hue comes from the glow of ionized oxygen (a relatively light element) heated by newborn stars.

Green peas have fascinated astronomers because, even though they have been observed only in the nearby Universe, their properties resemble those expected for early galaxies. The first galaxies, for example, are thought to contain only light elements because heavy elements need time to form. Sangeeta Malhotra, an astronomer at Goddard Space Flight Center, joked at the AAS meeting on 12 January that they are "Peter Pan galaxies" because they haven't matured.

Early doppelgängers

This is why Rhoads, Malhotra and their colleagues were excited by the spectra of the trio of galaxies analysed by JWST — the galaxies are from the early cosmos but resemble nearby green peas. "We knew they looked like green-pea spectra the minute we saw them," Rhoads told *Nature*.

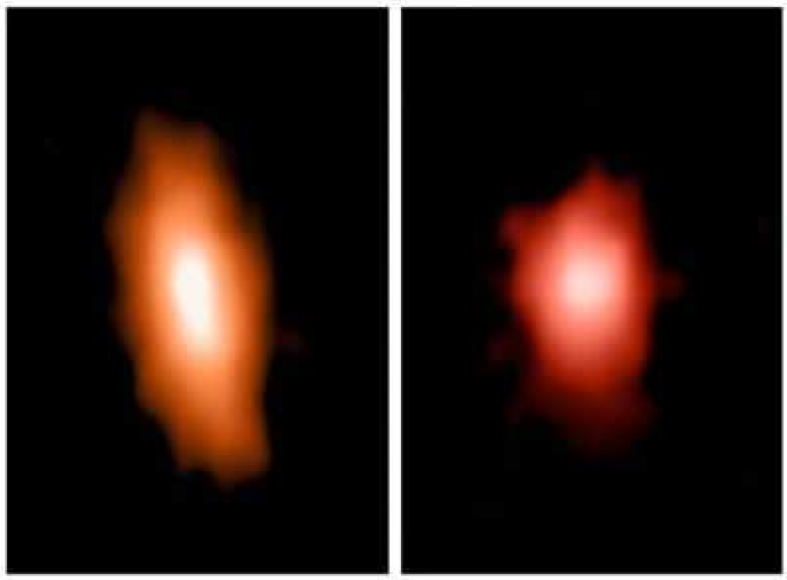


[Landmark Webb telescope releases first science image — astronomers are in awe]

Like green peas, the trio are small and have strong emission spectra — meaning they also churn out stars quickly. Similarly, they are low in elements heavier than hydrogen and helium. In fact, the most distant of the three contains roughly 2% as much oxygen as the Milky Way. That's possibly the lowest amount of oxygen seen in a galaxy yet.

Thuan says he was "amazed" to see the similarity between the spectra that JWST took of these galaxies and those of green peas, which he has long studied.

There is little doubt that they are the same type of object. Before JWST, researchers had measured green peas from ten billion years after the Big Bang, says Daniel Schaerer, an astronomer at the University of Geneva in Switzerland. "And now suddenly we can do the same thing — but just 700 million years after the Big Bang. It's completely mind-boggling."



[Four revelations from the Webb telescope about distant galaxies]

Another reason for excitement dates back to 2016, when astronomers including Thuan reported a green pea sending droves of ionizing photons into the intergalactic medium[3]. Most galaxies observed in the nearby Universe don't do this. With so many stars in such a small volume, green peas might be able to punch channels through the interstellar medium, creating an escape route for energetic radiation.

That could mean that objects like the trio from the early Universe produced the energy required to free the cosmos from its dark age. For Thuan, the results are convincing. "Now I really do think that these star-forming dwarf galaxies are the agent of reionization," he says.

Nature **613**, 425-426 (2023)

doi: https://doi.org/10.1038/d41586-023-00064-7

References

1. Rhoads, J. E. *et al. Astrophys. J. Lett.* **942**, L14 (2023).

2. Cardamone, C. *et al. Mon. Not. R. Astron. Soc.* **399**, 1191–1205 (2009).

3. Izotov, Y. I. *et al. Nature* **529**, 178–180 (2016).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00064-7

- NEWS
- 11 January 2023

2022 was a record year for space launches

180 rockets lifted off successfully, with SpaceX driving the pace.

- Alexandra Witze



NASA’s new Space Launch System rocket, carrying the Orion capsule, lifts off on 16 November 2022 as part of the Artemis I mission. Credit: Steve Nesius/Reuters

2022 was a record year for space with 180 successful rocket launches to orbit — the most ever, and 44 more than in 2021. The launches were dominated by rockets from US company SpaceX and from the Chinese government and businesses.

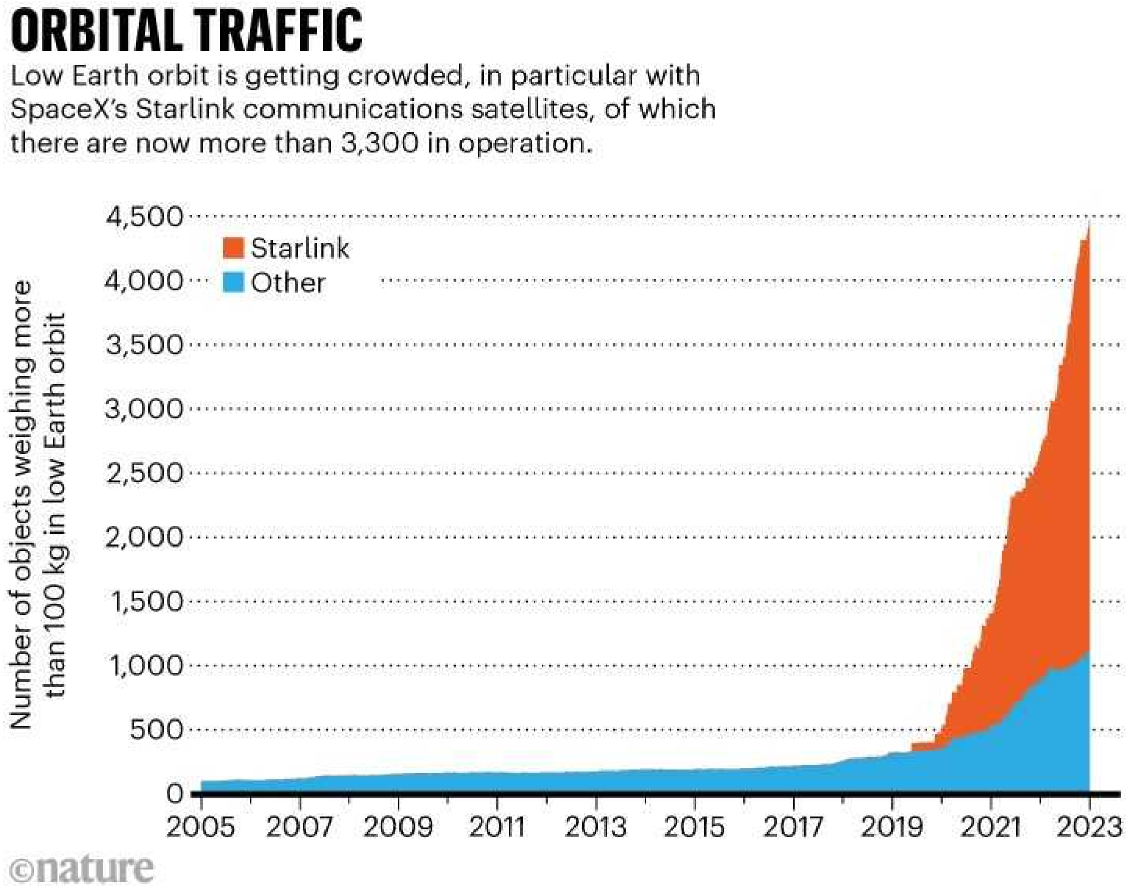
“Space traffic is shooting up, pun intended,” says Jonathan McDowell, an astrophysicist at the Center for Astrophysics Harvard & Smithsonian in Cambridge, Massachusetts, who maintains space-flight databases and this month published his analysis of launches in 2022.



The quest to conquer Earth’s space junk problem

SpaceX, headquartered in Hawthorne, California, sent one of its Falcon rockets into orbit once every six days, on average, throughout the year. Its total of 61 successful launches is tied with a 1980 record, for Soviet R-7 rockets.

Most of the SpaceX launches carried commercial payloads, including the company’s own network of Starlink communications satellites. There are now more than 3,300 operational Starlinks, by far the largest satellite constellation ever (see ‘Orbital traffic’). The growing congestion in space means that, between late 2020 and late 2022, SpaceX has had to move its Starlinks out of the way more than 26,000 times to prevent collisions with other objects in space.



Data from [Jonathan McDowell](#)

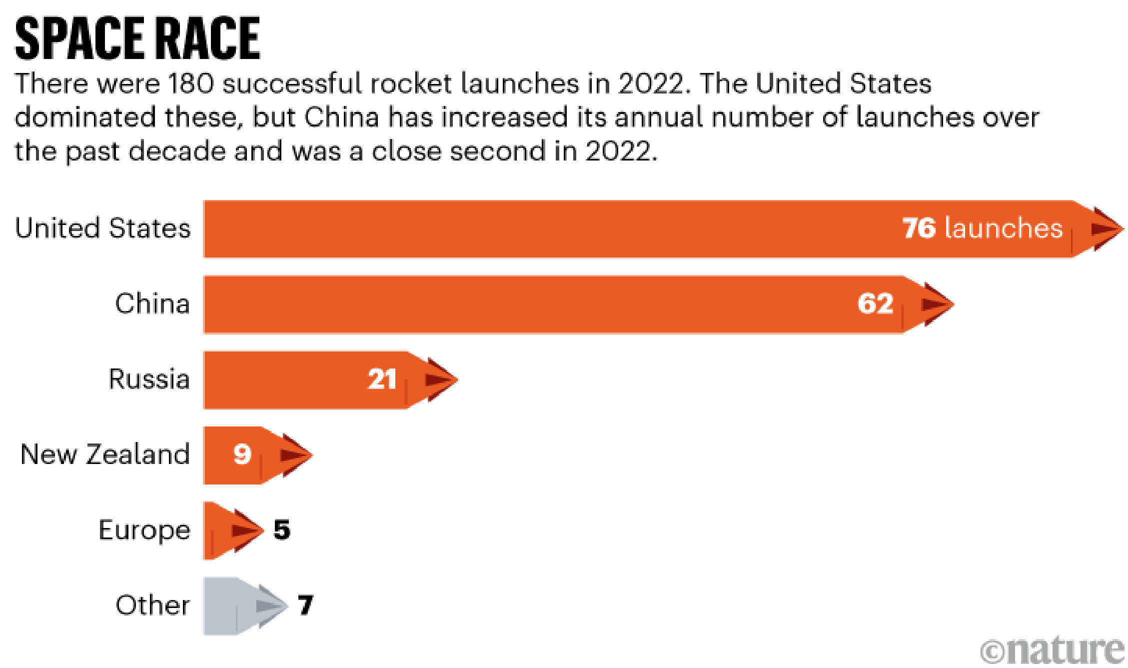
In addition to SpaceX, other US rocket providers accomplished 17 launches in 2022, according to McDowell's analysis. They included the November launch of NASA's new heavy-lift rocket, the Space Launch System, which is meant to ferry astronauts to the Moon in the coming years.

Growing space presence

China conducted 62 successful launches, 9 more than in 2021 (see 'Space race'). Many were government launches, but a quickly growing fraction belongs to commercial rocket providers. Overall, China's launch rate in 2022 was almost triple that of Russia. "China is replacing Russia as the number-two space power," McDowell says.

In October, China launched its solar observatory, nicknamed Kuafu-1. It also sent up the third and final major module of its orbiting space station, Tiangong, which is now complete.

Other scientific launches of the year included NASA's water-mapping Surface Water and Ocean Topography satellite, a private lunar lander from the Japanese company ispace and South Korea's first Moon mission.



Data from Jonathan McDowell

Europe's launch total plummeted from 15 successful launches in 2021 to only 5 last year. The European Space Agency stopped launching Russian-made Soyuz rockets after Russia invaded Ukraine in February 2022.

New Zealand, by contrast, saw a record-setting nine launches in 2022. The US company Rocket Lab sends commercial payloads into orbit from Mahia, New Zealand.

The coming year could break more records than the last. SpaceX has said it hopes to launch 100 times in 2023 — missions will include the first orbital flight of its enormous Starship vehicle, which is meant to eventually carry humans to the Moon and Mars.

Nature **613**, 426 (2023)

doi: https://doi.org/10.1038/d41586-023-00048-7

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00048-7

| Section menu | Main menu |

- NEWS FEATURE
- 18 January 2023

How your first brush with COVID warps your immunity

The immune system responds more strongly to the strain of a virus that it first met, weakening response to other strains. Can this ‘imprinting’ be overcome?

- Rachel Brazil

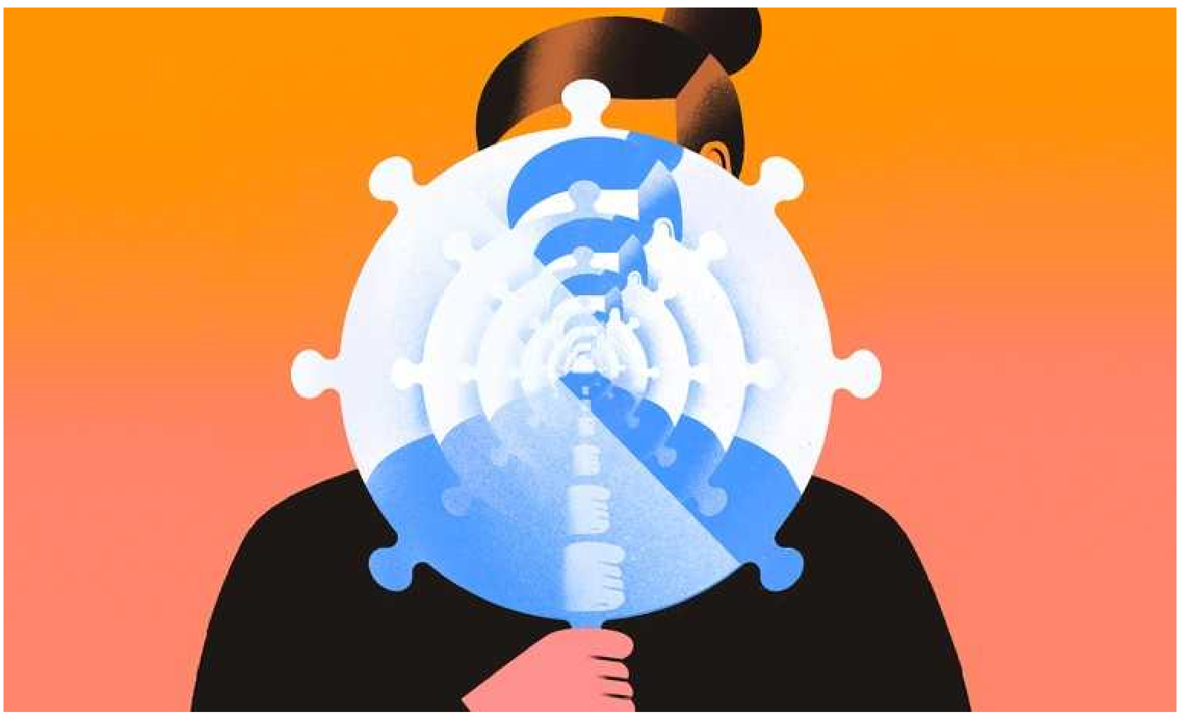


Illustration by Kasia Bojanowska

During the summer of 2022, with the Omicron coronavirus variant running rampant, friends and relatives of immunologist Bob Seder kept asking him if

they should postpone their COVID-19 boosters and wait for the new Omicron-tailored vaccine to become available. He told them not to delay.

Seder, acting chief of the Vaccine Immunology Program at the US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, suspected that the effectiveness of a new booster would be blunted by a quirk of the immune system known as imprinting — the tendency for someone's initial exposure to a virus to bias their immune response when they meet the same virus again.

Imprinting was first observed decades ago, in people with influenza. Their immune systems responded to a new circulating strain by producing antibodies tailored to their first flu encounter. In some cases, this led to a poorer ability to fight off the new strain.

The phenomenon can explain some observations from the past, such as the surprisingly high mortality among young adults during the 1918 influenza pandemic. Members of the older generation, exposed in their youth to a flu strain that closely matched the deadly H1N1 pandemic strain, had a more robust immune response than did younger adults, whose first exposure was to a mismatched strain.

A crop of studies is now showing how imprinting is shaping people's response to SARS-CoV-2. For example, those infected with the earliest strain or with the subsequent Alpha or Beta strains mount varying immune responses to a later Omicron infection, depending on the strain to which they were first exposed.

Furthermore, even exposure to Omicron itself doesn't seem to help update the imprinted response of people previously infected with an older strain, which might explain why they can be reinfected.



The ghost of influenza past and the hunt for a universal vaccine

It’s now relatively easy to update mRNA vaccines to match a new strain, but imprinting suggests that these tailored vaccines might not significantly improve protection against infection. And although they are clearly able to prevent serious illness, this puts a dampener on the hope that variant-tailored boosters will markedly reduce transmission of the virus.

Still, researchers agree that variant-tailored boosters are worth getting because they still provide some immunity, and that imprinting will not make COVID-19 more severe than it would be in someone with no previous exposure. “You’re better off having some immunity, no matter what it is,” says Katie Gostic, an evolutionary biologist at the University of Chicago, Illinois.

What’s more, there are hints that, in some people at least, the immune system can adapt, raising the possibility of improving immune responses.

Immune memory

Imprinting equips the immune system with a memory of an invader that helps it prepare to do battle again. The key players are memory B cells, which are generated in lymph nodes during the body’s first exposure to a

virus. These cells then keep watch in the bloodstream for the same foe, ready to develop into plasma cells that then churn out antibodies. The snag comes when the immune system encounters a similar, but not identical, strain of a virus. In this case, rather than generate new, or ‘naive’, B cells to produce tailored antibodies, the memory-B-cell response kicks in. This often leads to the production of antibodies that bind to features found in both the old and new strains, known as cross-reactive antibodies. They might offer some protection but are not a perfect fit to the new strain.

Imprinting was first observed in 1947 by Jonas Salk and Thomas Francis, the developers of the first flu vaccine, together with another scientist, Joseph Quilligan[1]. They found that people who had previously had flu, and were then vaccinated against the current circulating strain, produced antibodies against the first strain they had encountered. Francis gave the phenomenon the tongue-in-cheek name ‘original antigenic sin’, although today most researchers prefer to call it imprinting.

Researchers have only recently demonstrated how strongly this process can influence immunity. In 2016, Gostic published an epidemiological study[2] that proved imprinting was more than just a curiosity. She analysed data from two pandemics of avian influenza A, caused by the viruses H5N1 in 2009 and H7N9 in 2013. Both strains share some characteristics with seasonal-flu strains, but they come from opposite sides of the flu evolutionary tree. “We saw this really remarkably clear pattern that you seem to be much more susceptible, at least to severe infection, if you had been imprinted in childhood to a mismatched subtype,” says Gostic.



US soldiers bound for France march through Seattle wearing masks in the midst of the influenza pandemic in 1918.Credit: Science History Images/Alamy

For SARS-CoV-2, too, “your infection history, and your vaccination history in combination, are imprinting your subsequent immune response when you see the live virus”, says immunologist Rosemary Boyton at Imperial College London. Boyton and her colleagues reached this conclusion by studying immunity in a large group of health-care workers at several London hospitals.

Their first study[3], conducted before Omicron had emerged, looked at the responses of certain facets of the immune system — including antibodies and B and T cells — in individuals who had received two vaccines. Some of them had been infected before their jabs and others after. The researchers found that a person’s immunity to infection by subsequent strains depended strongly on their previous infections or vaccinations. “Some combinations happen to offer better future- proofing against an oncoming variant than others,” says Boyton.

Last June, she and her colleagues published a follow-up study[4] looking at immunity in a group of people who had contracted an Omicron infection after triple vaccination, but who had various COVID-19 infection histories. Again, they saw a variety of responses indicating imprinting from previous exposure. For example, even in people whose first COVID-19 infection was with Omicron, the antibodies were a better match to the original strain — against which they had been vaccinated — and to the older Alpha and Delta strains.

For those who had been previously infected with the original strain, and were then vaccinated (with a shot designed to fight that strain), the subsequent Omicron infection didn't boost their ability to create Omicron-adapted antibodies at all. This is a clear sign of imprinting, Boyton says, and probably explains why Omicron reinfections are so common — although, for most people, even an imprinted response is enough to stop serious illness.



Do childhood colds help the body respond to COVID?

This idea is backed up by a study[5] published last month by Yunlong Cao at Peking University in Beijing. He found that people who had been vaccinated with the original strain and then contracted an Omicron infection produced antibodies that were mainly cross-reactive to both strains, but rarely specific to Omicron itself. By contrast, those without a previous vaccination made

antibodies that specifically matched Omicron. Cao says that, as the two strains gradually diverged from each other, the proportion of antibodies that could neutralize the second strain decreased. He expects this effect to be even more pronounced with the newer Omicron subvariants, such as XBB.

But the immune system has some tricks that could help to counteract the effect of imprinting, says Laura Walker, chief scientific officer and co-founder of antibody-focused pharmaceutical company Invivyd, based in Waltham, Massachusetts. Memory B cells can mutate, to some extent, when exposed to a new strain, producing better-matched antibodies in a process known as affinity maturation. Walker and her colleagues tracked antibody responses in mRNA-vaccinated individuals for up to six months after they had contracted Omicron, and found that at least a subset of the B cells was altered so that they started to produce antibodies matching Omicron[6]. Boyton agrees that affinity maturation might lead immune repertoires to change over time, but the extent to which this is happening after multiple vaccinations is not yet clear.

The transformation occurs in transient structures in the lymph nodes and bone marrow called germinal centres. "You can think of it as a boot camp, where our immune cells get trained to become even better," says Ali Ellebedy, an immunologist at Washington University in St. Louis, Missouri. Those B cells then go on to produce more-effective antibodies.



A health worker applies a sticking plaster after giving a man a booster shot of the Sinovac vaccine in Manila, Philippines this month.Credit: Lisa Marie David/Reuters

Last September, Ellebedy reported[7] the results of a study in which his team collected B cells from the germinal centres of volunteers who had received a booster vaccine targeting either the original SARS-CoV-2 strain or Omicron.

In both cases, even when the original strain was not present in the booster, antibodies recognizing the original strain were dominant. But, for the Omicron booster, Ellebedy says, “we were able to detect a very small fraction of cells that responded specifically to Omicron”. This suggests that imprinting does not completely quash responses to new strains — although the team didn’t see this effect in all individuals. Key questions are why that is and how to encourage this new response.

Although affinity maturation helps to align existing B cells with a new foe, researchers have also looked at whether the immune system can deploy entirely new B cells when faced with a new infection. Immunologist Gabriel Victora at the Rockefeller University in New York City developed a

technique that tracks cells and their descendants in mice, to figure out which antibodies came from which B cells.

His results[8], published this month, showed that when mice previously vaccinated with the original SARS-CoV-2 strain were boosted with that strain, more than 90% of the antibodies produced were derived from pre-existing B cells. But boosting the animals with the Omicron strain changed the picture, says Victora: 25–50% of the antibodies came from new B cells, and were better at neutralizing Omicron than were the older, recycled antibodies.

Walker says that it's unclear how this translates into longer-term protection. For that to occur, the new B cells would need to mature into antibody-producing plasma cells in the bone marrow, and it's not clear whether the immune system retains these newer plasma cells.

Imprinting impact

Although antibody studies reveal the telltale signature of imprinting, Gostic says, there is little evidence that those signatures affect people's susceptibility to illness. Even if it lessens protection, says Boyton, "there's no evidence that it causes harm, makes a worse immune response, or makes you more sick" compared with people who are unvaccinated or haven't been infected.

Gostic prefers to look on the bright side: rather than imprinting lessening the overall response, she says "you have a sort of superpower if you have been exposed to something that's matched".

But imprinting does seem to have dashed the hope that the introduction of variant-targeted mRNA boosters will provide greater protection against infection than sticking with the original vaccine. Boosters introduced in Europe last September targeted the original strain and the BA.1 Omicron variant; the United States has rolled out boosters aimed at the BA.5 Omicron variant. The vaccines certainly boost antibody levels, but the antibodies produced are not Omicron specific and are unlikely to offer significantly increased protection against Omicron infection[9,10].



China and India approve nasal COVID vaccines — are they a game changer?

So what does this mean for our current vaccines? Boyton says that they are “brilliant” in their ability to protect against serious illness. But, she says, now that most people are protected, scientists should focus on finding vaccines that can overcome imprinting, to halt the spread of the virus, not just the severity of disease. “Now we’re in a slightly different place, we’ve got to think slightly differently.”

Seder agrees that vaccines will need to change if they are to limit infection and transmission, rather than just prevent fatalities. He says that one approach would be to use live vaccines, which would persist in the body for 5–10 days and might produce a more effective response. But live vaccines pose greater risks, particularly for vulnerable people, owing to the dangers of even a weakened virus multiplying.

Instead, Seder is looking at nasal vaccines, which he thinks could be more effective against variants than injected vaccines. Spraying a vaccine directly into the nose could induce mucosal immunity — an immune response in the cells that make up the lining of the respiratory system and other mucous membranes. In a natural infection, the mucous membranes are the first barrier a virus will encounter. The antibody response here is prolific and designed to thwart virus entry. The response might still carry some imprint

of past exposure, but its sheer strength could create better protection — even preventing infection and transmission. A study[11] published last year showed that the antibody response produced by injected vaccines bypasses the antibody cascade that mucosal immunity induces, whereas a breakthrough infection does induce this response.

Seder is now comparing nasal and injected mRNA vaccines in animals to see whether they can improve the response to Omicron. Data on how these vaccines perform in humans are sparse, and mixed. In October, a phase I clinical trial of a vaccine developed by Astrazeneca and the University of Oxford, UK, administered as a nasal spray, reported mucosal antibody responses in only a minority of participants, and weaker systemic responses than were obtained with intramuscular vaccination[12].

Another approach is the use of adjuvants — ingredients added to vaccines that help to boost the immune response. Adjuvants have been shown to alleviate imprinting in vaccination against flu[13]. So far, their effect on imprinting in COVID-19 has not been tested.

The ideal immune response is strong and broad — both for disarming imprinting and for fighting a wider range of viruses and variants, says Boyton. A vaccine against SARS-CoV-2 and all its relatives — a pan-coronavirus vaccine — would induce a wider variety of antibodies that target multiple parts of the virus. Such a vaccine could prevent the virus from mutating enough to escape the immune system, and might ultimately be the key to controlling future pandemics. The hunt for such a vaccine is ongoing.

Imprinting is often presented as a problem, but it’s a fundamental part of immune memory that delivers a lightning-fast response to a viral invader, without having to start from scratch. “To me,” says Victora, “it means that the immune system is quite smart at covering its bases.”

Nature **613**, 428-430 (2023)

doi: https://doi.org/10.1038/d41586-023-00086-1

References

1. Francis, T. Jr, Salk, J. E. & Quilligan, J. J. Jr *Am. J. Public Health* **37**, 1013–1016 (1947).

2. Gostic, K. M., Ambrose, M., Worobey, M. & Lloyd-Smith, J. O. *Science* **354**, 722–726 (2016).

3. Reynolds, C. J. *et al. Science* **375**, 183–192 (2022).

4. Reynolds, C. J. *et al. Science* **377**, eabq1841 (2022).

5. Cao, Y. *et al. Nature* https://doi.org/10.1038/s41586-022-05644-7 (2022).

6. Kaku, C. I. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.09.21.508922 (2022).

7. Alsoussi, W. B. *et al.* Preprint at bioRxiv https://doi.org//10.1101/2022.09.22.509040 (2022).

8. Schiepers, A. *et al. Nature* https://doi.org/10.1038/s41586-023-05715-3 (2023).

9. Wang, Q. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.10.22.513349 (2022).

10. Collier, A. Y. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.10.24.513619 (2022).

11. Park, Y.-J. *et al. Science* **378**, 619–627 (2022).

12. Madhavan, M. *et al. eBioMedicine* **85**, 104298 (2022).

13. Kim, J. H., Davis, W. G., Sambhara, S. & Jacob, J. *Proc. Natl Acad. Sci. USA* **109**, 13751–13756 (2012).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00086-1

Books & Arts

- **How mathematics stopped being defined by reality — and started to invent new ones** [16 Jan 2023]
  • Abstraction is a powerful mathematical technique that has influenced everything from quantum theories of gravity to nuclear deterrence theory. But has it gone too far?
- **How commercial rivalry is ruining space, and more: Books in brief** [16 Dec 2022]
  • Andrew Robinson reviews five of the best science picks.

- BOOK REVIEW
- 16 January 2023

How mathematics stopped being defined by reality — and started to invent new ones

Abstraction is a powerful mathematical technique that has influenced everything from quantum theories of gravity to nuclear deterrence theory. But has it gone too far?

- Davide Castelvecchi



Euclid’s *Elements* provided the basis for mathematics.Credit: Tim Ockenden/PA/Alamy

Axiomatics: Mathematical Thought and High Modernism *Alma Steingart* Univ. Chicago Press (2022)

The work of mathematicians from centuries or even millennia ago speaks to their living peers in ways that practitioners of other disciplines must find baffling. Euclid's proof that the list of prime numbers never ends is just as elegant and clear now as it was in around 300 bc, when it appeared in his book *Elements*.

Yet mathematics has undergone tremendous changes, especially during the twentieth century, when it pushed ever deeper into the realm of abstraction. This upheaval even involved a redefinition of the definition itself, as Alma Steingart explains in *Axiomatics*.

A historian of science, Steingart sees this revolution as central to the modernist movements that dominated the mid-twentieth century in the arts and social sciences, particularly in the United States. Mathematicians' push for abstraction was mirrored by — and often directly triggered — parallel trends in economics, sociology, psychology and political science. Steingart quotes some scientists who saw their liberation from merely explaining the natural world as analogous to how abstract expressionism freed painting from the shackles of reality.



[Maths classification is getting a revision]

The mathematics encapsulated in Euclid's *Elements* started from foundational truths called axioms and built up statements through logical deductions. To Euclid, axioms such as 'any straight segment can be extended to an infinite line' were self-evident because they were rooted in physical reality.

For centuries, Western culture revered *Elements* as the paragon of intellectual rigour. In the seventeenth century, political theorist Thomas Hobbes referred to the book as 'a structural model of a stable society' and philosopher Baruch Spinoza adopted its structure of axioms and deductions in his treatise on ethics. The Catholic Church, meanwhile, held Euclid's work as dogma, a way to anchor itself amid what it considered the anarchy brought about by the Protestant Reformation (as well as to protect it from the revolutionary threat of 'infinitesimally small' numbers introduced in calculus, as Amir Alexander explains in his intriguing 2014 book *Infinitesimal*).

Diverging worlds

Modern mathematics turns Euclid's principle of axioms derived from physical reality on its head. Mathematicians might follow their creative impulses to come up with any set of axioms they wish — inventing, for example, non-Euclidean geometries in which parallel lines converge or diverge.

In this conception, creating a mathematical theory is similar to setting the rules of a game such as chess, in which the names, shapes and roles of the pieces are purely a matter of convention. Just as someone can begin with the rules of chess and a given configuration of the board and arrive at a prediction of 'checkmate in five moves', a mathematician can start from axioms and go through a sequence of logical steps to prove the truth of a theorem, without worrying what reality it represents.

BOOK III.

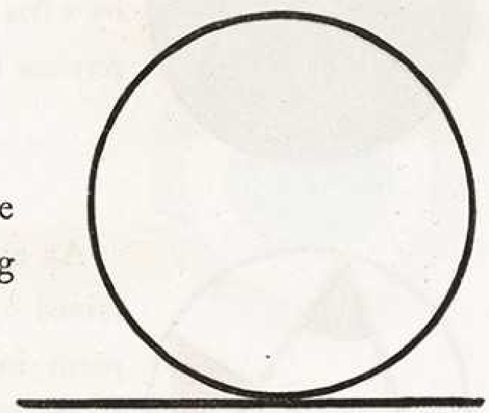
DEFINITIONS.

I.

EQUAL circles are thoſe whoſe diameters are equal.

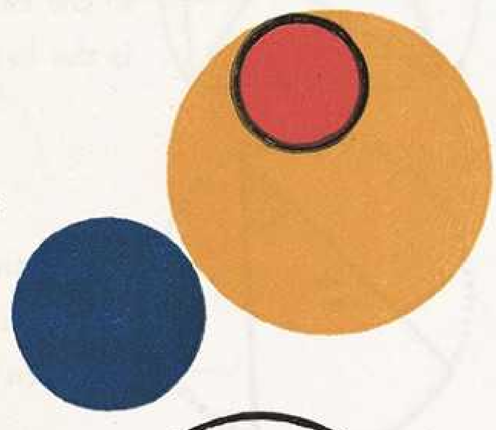
II.

A right line is said to touch a circle when it meets the circle, and being produced does not cut it.



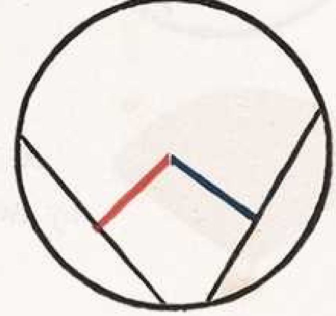
III.

Circles are ſaid to touch one another which meet but do not cut one another.



IV.

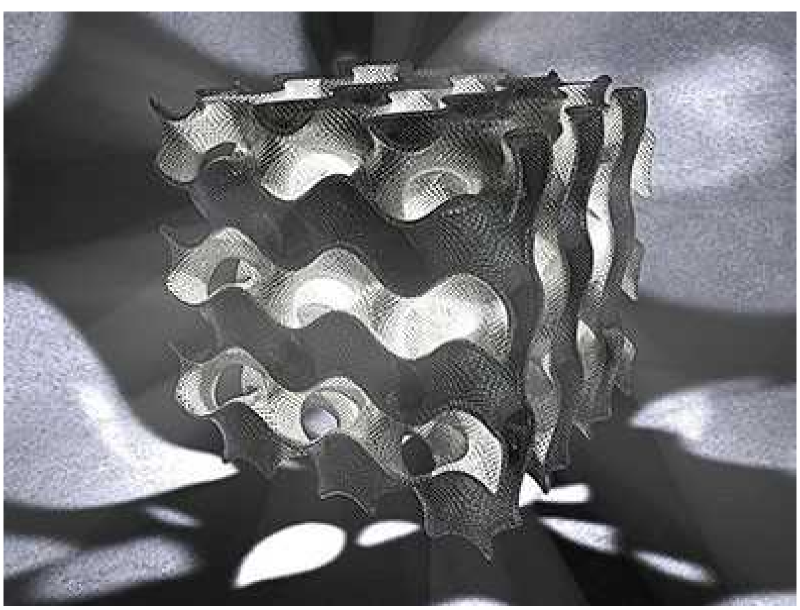
Right lines are ſaid to be equally diſtant from the centre of a circle when the perpendiculars drawn to them from the centre are equal.



Euclid's geometry was rooted in reality — but modern mathematics is less constrained.Credit: SSPL/Getty

To the mathematical-theory builder, abstraction is not a destination, but a journey. As Steingart puts it, 'abstract' is not an adjective but a verb: 'to abstract'. In the 1930s, owing largely to the influence of German mathematician Emmy Noether, mathematicians began to construct axiomatic systems that were increasingly abstract and general. This revealed familiar objects such as numbers, card shuffles and geometrical symmetries to be special cases of the same concept.

The trend towards abstraction and generalization is often associated with a school of mathematics that blossomed in France after the Second World War. But, as Steingart shows, it took root in the 1930s in the United States and came to define the country's mid-century mathematical culture. Steingart exemplifies the trend with the story of *Foundations of Algebraic Topology*, a 1952 book by US mathematicians Samuel Eilenberg and Norman Steenrod. It dealt with various calculation techniques to distinguish between geometric shapes, but the authors introduced the subject backwards, claiming that students should first familiarize themselves with highly technical algebraic tools and only later learn their relevance to shapes, or why the tools existed in the first place.



The strange topology that is reshaping physics

A central thesis in *Axiomatics* is that US mathematicians saw the process of abstraction and generalization not as antithetical to the real world, but as crucial to how maths — or, rather, mathematical thinking — can be applied to practical problems. ‘Mathematizing’ a problem did not mean to measure and compute, but to reveal a hidden skeleton of conceptual relationships: to formulate the underlying idea in abstract mathematical language.

Often, this meant not applying existing mathematical concepts, but inventing new ones. An influential example was the 1944 book *Theory of Games and Economic Behavior*, which established modern game theory and included very non-playful implications for strategies of nuclear defence. By retracing an impressive web of connections, Steingart shows how the authors, the US mathematician John von Neumann and the German economist Oskar Morgenstern, not only helped to define the relationship between science and politics during the cold war, but also provided a model for further attempts at mathematization in the social sciences.

Taken to excess

Ironically, even as some mathematicians argued that abstract thinking was the key to applying maths to other disciplines — and suggested that even the most abstract maths was worthy of public funding — most mathematicians in academia seemed singularly uninterested in getting involved. For decades, much of the progress in applied mathematics ended up taking place not in universities, but in think tanks and industry laboratories, or in newly established departments devoted to fields such as computer science or statistics.

Eventually, an excess of abstraction caught up with mathematicians. Steingart sketches how the latter parts of the twentieth century saw a turning of the tide: she picks the example of William Thurston, an enormously influential topologist who delighted in making his complex geometric constructions feel physically real.

One major development she does not touch on is a renewed cross-fertilization with theoretical physics in the late twentieth century. This includes the application of topology to innovations such as ‘topological’

materials — which could even form a basis for super-powerful quantum computers — and the development of string theory, which might not have given physicists their long-sought theory of everything, but has inspired many a maths PhD thesis. And there is some anecdotal evidence that in the past decade or so, the barriers between pure and applied mathematics have started to drop: it is not uncommon now to see researchers from the most abstract reaches of the field 'getting their hands dirty' with applications such as data analysis.

Meanwhile, Eilenberg and Steenrod's approach to pedagogy came to be seen as a cautionary tale, with their techniques, although still widely used, affectionately being called 'abstract nonsense'. But in another of the twists of fate often seen in the history of maths and physics, some physicists now consider abstract-nonsense techniques a promising approach to devising a quantum theory of gravity — perhaps delivering another route to the very real, if abstract, goal of a theory of everything.

Nature **613**, 431-432 (2023)

doi: https://doi.org/10.1038/d41586-023-00087-0

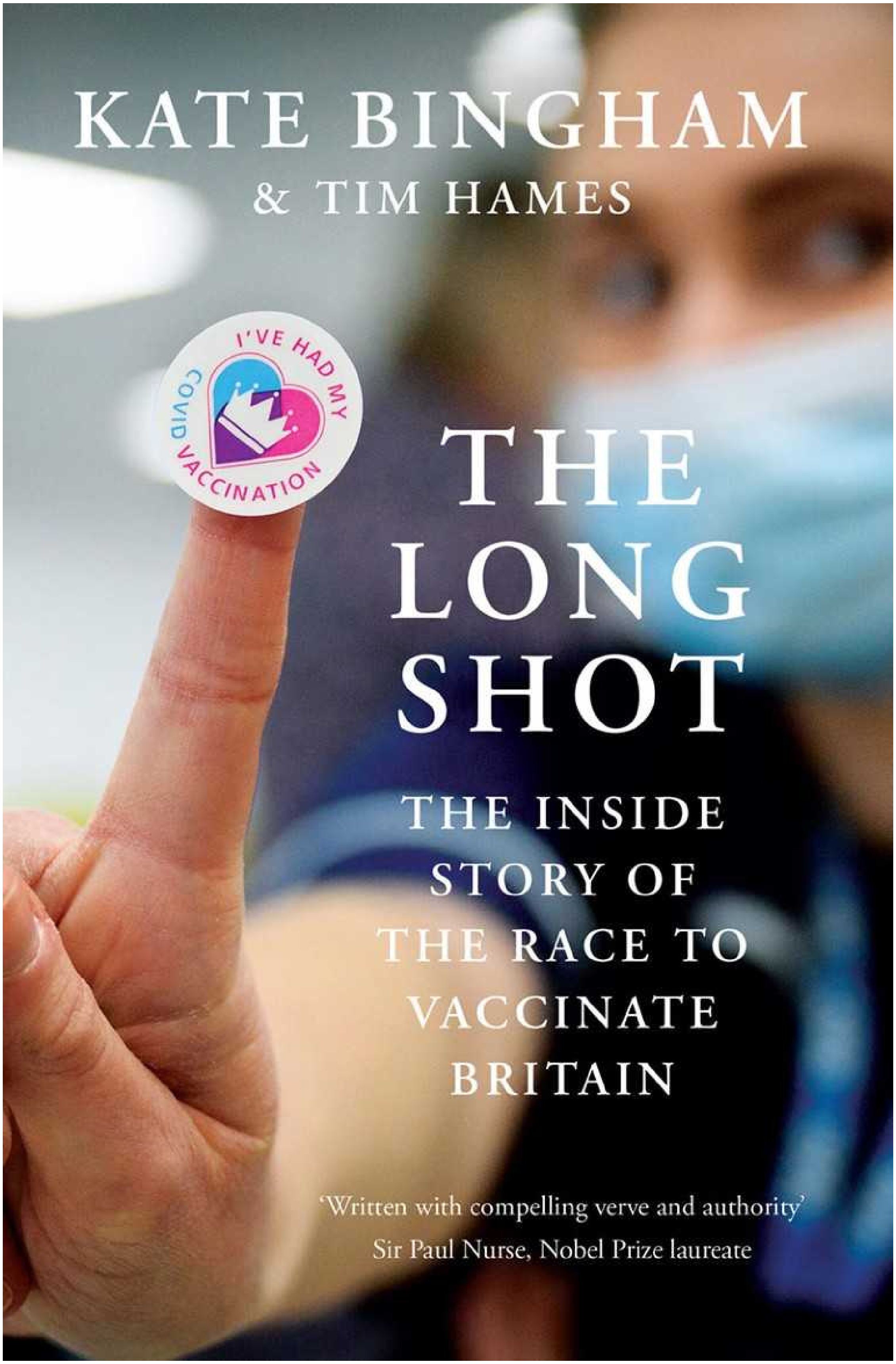
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00087-0

- BOOK REVIEW
- 16 December 2022

How commercial rivalry is ruining space, and more: Books in brief

Andrew Robinson reviews five of the best science picks.

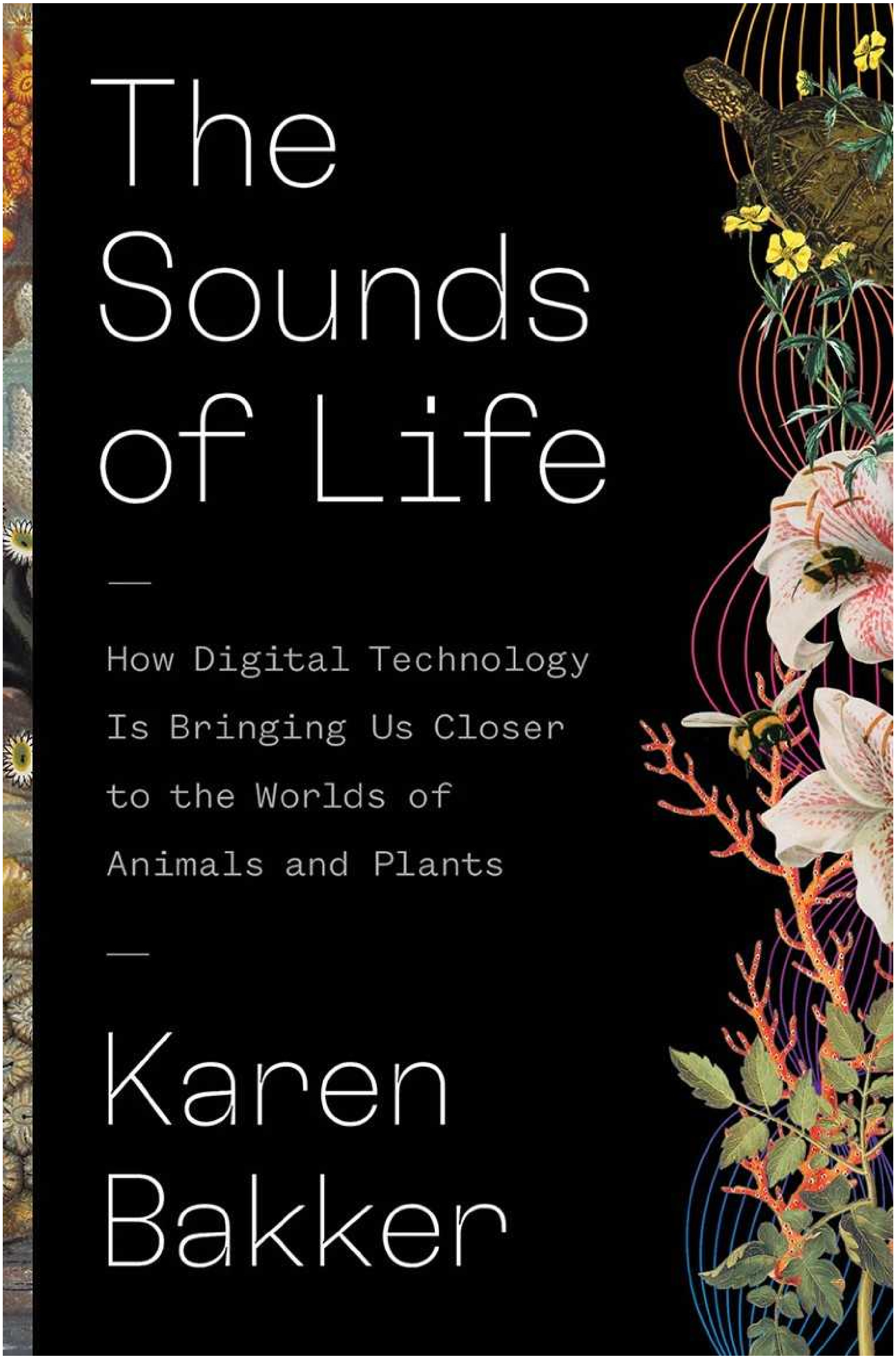
- Andrew Robinson


KATE BINGHAM
& TIM HAMES
I'VE HAD MY COVID VACCINATION
THE LONG SHOT
THE INSIDE STORY OF THE RACE TO VACCINATE BRITAIN
'Written with compelling verve and authority'
Sir Paul Nurse, Nobel Prize laureate

The Long Shot

Kate Bingham and Tim Hames *Oneworld* (2022)

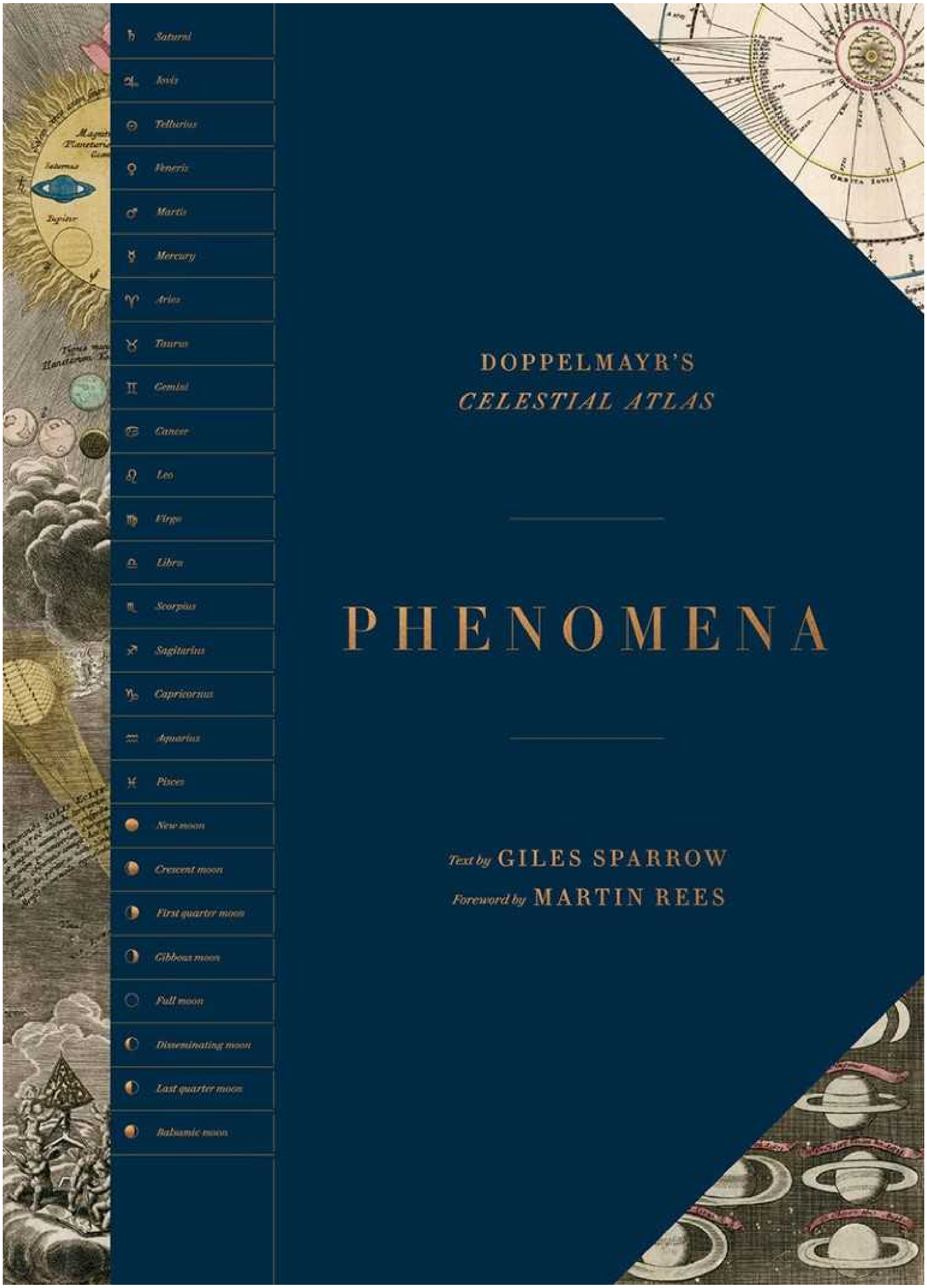
In April 2020, the UK government decided to expedite the introduction and deployment of a COVID-19 vaccine. Eight months later, the first recipient was jabbed. Credit goes to the Vaccine Taskforce led by Kate Bingham, a biochemistry graduate and health venture capitalist, working from home during lockdown. She was acutely aware of most vaccines' failure, the risk of viral mutation and her lack of government training. Her frank and fascinating book, written with journalist Tim Hames, details the inside story of this "longest of all long shots".


The Sounds of Life
How Digital Technology Is Bringing Us Closer to the Worlds of Animals and Plants
Karen Bakker

The Sounds of Life

Karen Bakker *Princeton Univ. Press* (2022)

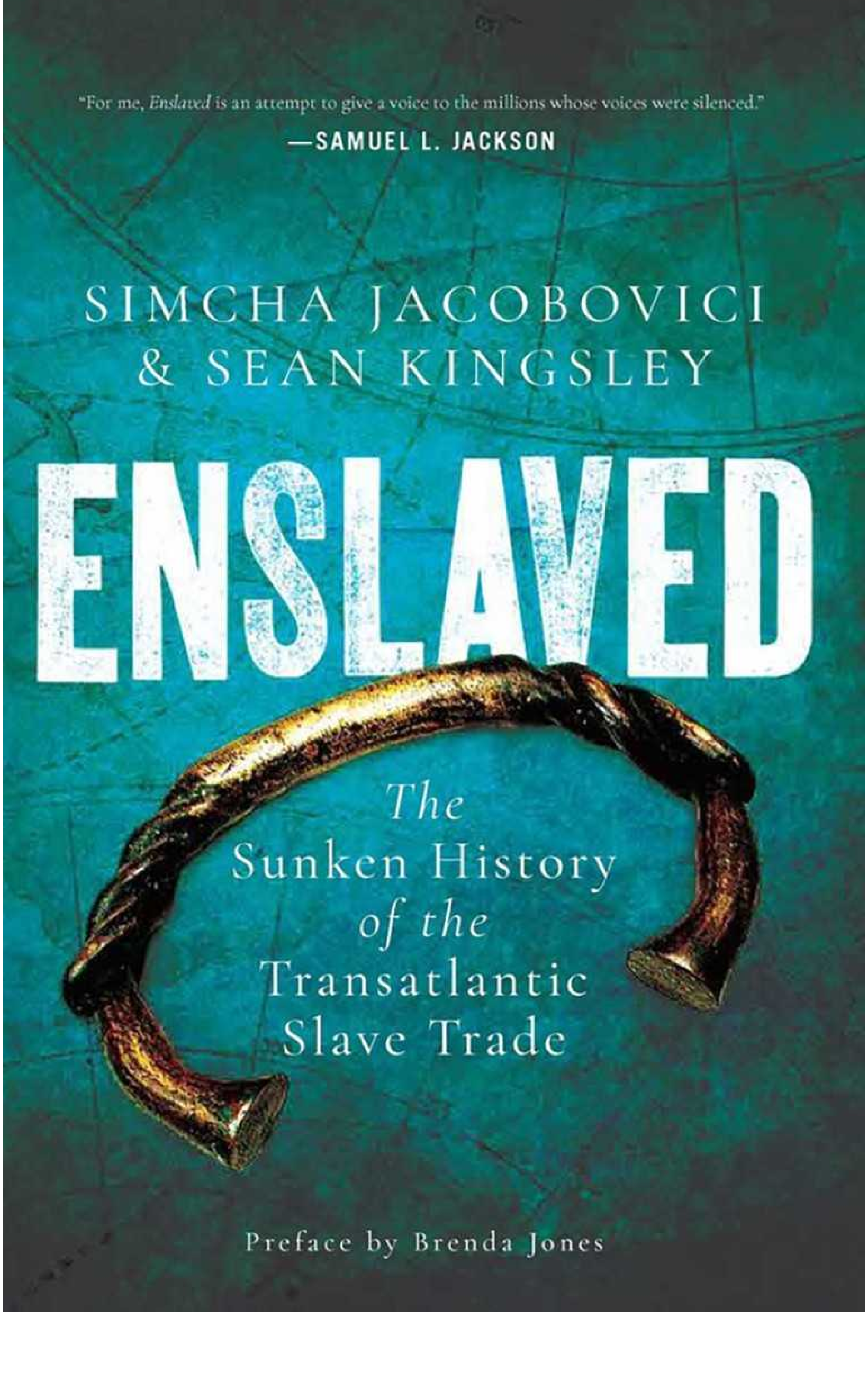
Nature is full of sounds that humans cannot hear. The raised-tail mating display of male peacocks involves powerful infrasound; bats and toothed whales emit beams of ultrasound and navigate using the echoes that come back. In this beautifully written study, Karen Bakker, a tech entrepreneur and academic, compares digital technology that can reveal these sounds with the microscope's effect on vision. By extending our hearing, the technology allows us to encounter "new soundscapes around the world and across the Tree of Life".


DOPPELMAYR'S
CELESTIAL ATLAS
PHENOMENA
Text by GILES SPARROW
Foreword by MARTIN REES

Phaenomena

Giles Sparrow *Thames & Hudson* (2022)

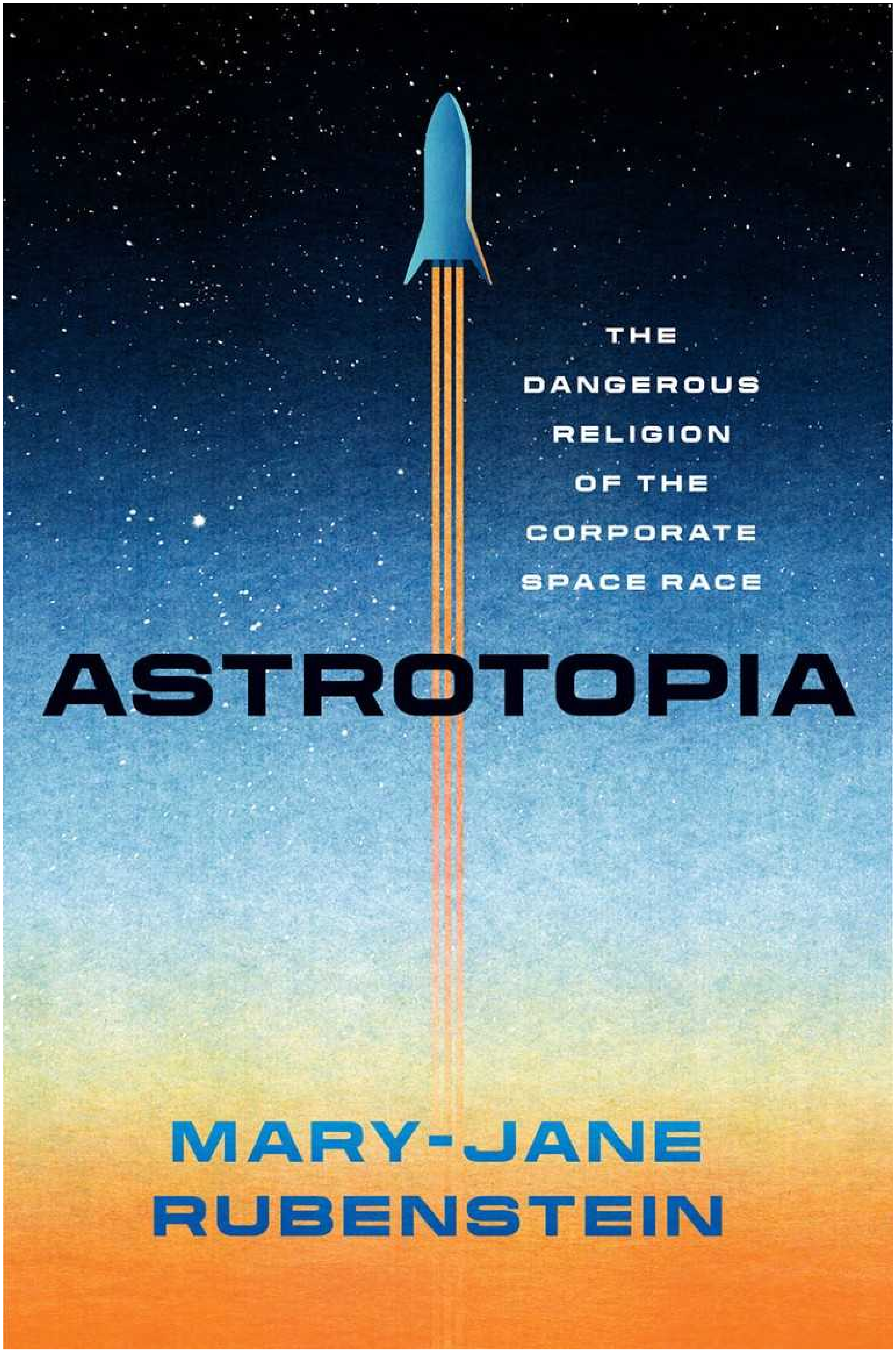
Mathematician, astronomer and cartographer Johann Doppelmayr created his largely forgotten celestial atlas in Germany in 1742. This magnificent book showcases its illustrations, with formidable explanatory text by astronomer Giles Sparrow and foreword by UK astronomer royal Martin Rees. The frontispiece depicted astronomers Ptolemy, Nicolaus Copernicus, Johannes Kepler and Tycho Brahe. Isaac Newton is absent; his 1687 concept of gravity was, Sparrow notes, too close to "occult forces" for many natural philosophers of the time.


"For me, *Enslaved* is an attempt to give a voice to the millions whose voices were silenced."
—SAMUEL L. JACKSON
SIMCHA JACOBOVICI
& SEAN KINGSLEY
ENSLAVED
The
Sunken History
of the
Transatlantic
Slave Trade
Preface by Brenda Jones

Enslaved

Simcha Jacobovici and Sean Kingsley *Pegasus* (2022)

During the 400-year transatlantic slave trade, 12 million Africans were forcibly shipped to the Americas, and almost 2 million died en route. Using mixed-gas diving and remotely operated vehicles, divers have investigated seven surviving deep-sea wrecks of ships used in the trade and two 'freedom ships' carrying fugitives from slavery, as described in a six-part series by film-maker Simcha Jacobovici. His book, written with marine archaeologist Sean Kingsley, vividly explores these disturbing sunken archaeological archives.


THE
DANGEROUS
RELIGION
OF THE
CORPORATE
SPACE RACE
ASTROTOPIA
MARY-JANE
RUBENSTEIN

Astrotopia

Mary-Jane Rubenstein *Univ. Chicago Press* (2022)

In the cold war, space exploration's wonders served a race between political systems. Today, argues religion and science scholar Mary-Jane Rubenstein, they are subject to commercial rivalry, notably between billionaires Jeff Bezos and Elon Musk, who are littering space and advocating it as a refuge from Earth's destruction. She argues that we must eschew such myopic, colonialist "astrotopia", and listen instead to a sort of "pantheistic mysticism" — valuing and learning from the cosmic environment — scorned by past imperialists.

Nature **613**, 432 (2023)

doi: https://doi.org/10.1038/d41586-022-04478-7

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-022-04478-7

Opinion

- **How games can make behavioural science better** [17 Jan 2023]
  • Wordle, Minecraft and Scrabble are played online by millions. Gamifying experiments can make behavioural research more inclusive, rigorous and reproducible — if it's done right.

- COMMENT
- 17 January 2023

How games can make behavioural science better

Wordle, Minecraft and Scrabble are played online by millions. Gamifying experiments can make behavioural research more inclusive, rigorous and reproducible — if it's done right.

- Bria Long ,
- Jan Simson,
- Andrés Buxó-Lugo,
- Duane G. Watson &
- …
- Samuel A. Mehr

Show authors



Researchers have added puzzles to the game Minecraft to help study behaviour.Credit: Matthew Tostevin/Reuters

When US cognitive scientist Joshua Hartshorne was investigating how people around the world learn English, he needed to get tens of thousands of people to take a language test. He designed 'Which English?', a grammar game that presented a series of tough word problems and then guessed where in the world the player learnt the language. Participants shared their results — whether accurate or not — on social media, creating a snowball effect for recruitment. The findings, based on data from almost 670,000 people, revealed that there is a 'critical period' for second-language learning that extends into adolescence[1].

This sort of 'gamification' is becoming a powerful research tool across fields that study humans, including psychology, neuroscience, economics and behavioural economics. By making research fun, the approach can help experiments to reach thousands or millions of participants. For instance, experiments embedded in a video game demonstrated that the layout of the city where a child lives shapes their future navigational ability[2]. Data from a

digital word search showed that people who are skilled at the game do not necessarily give better advice to those trying to learn it[3]. And a dilemma game involving millions of people revealed that most individuals have reliable moral intuition[4].

Gamification can help to avoid the pitfalls of conventional laboratory-based experiments by allowing researchers to study diverse populations, to conduct more-sophisticated experiments and to observe human behaviour in naturalistic environments. It can improve statistical power and reproducibility, making research more robust. Technical advances are making gamification cheaper and more straightforward, and the COVID-19 pandemic has forced many labs to move their human experiments online. But despite these changes, most have not yet embraced the opportunities gamification affords.

To reach the full potential of this approach, researchers must dispel misconceptions, develop new gamification technologies, improve access to existing ones and apply the methods to productive research questions. We are researchers in psychology, linguistics, developmental science, data science and music who have run our own gamified experiments. We think it's time for science to get serious about games.

Games theory

Gamification motivates people to participate in experiments by incorporating point-scoring, competition, feedback about performance and the opportunity to learn about oneself through play (see 'Take part in gamified research'). In early forays in 2005 and 2008, cognitive neuropsychologist Laura Germine, now at Harvard Medical School in Boston, Massachusetts, adapted psychological assessments and placed them on her citizen-science website, TestMyBrain.org. Her work showed that self-selected samples can produce high-quality data even when participants are unsupervised and unpaid[5]. As gamification has developed, it has united developmental psychology with computer science, web development and user-experience research to create exciting, immersive encounters for participants.

Take part in gamified research

Gamification motivates people to participate in experiments through competition, fun and the opportunity to learn about ourselves. Here are some games to try out:

Which English? A viral grammar test that tries to identify which "world English" someone speaks, as part of language research.

Are you a Super-Listener? A citizen-science experiment in which participants try to detect musical harmonies in scrambled tones.

Glyph: An online gaming applet to explore the shape of letters in the world's writing systems.

Moral Machine: A platform for gathering a human perspective on moral decisions made by machine intelligence.

Visual Vocab: An online assessment of vocabulary knowledge across the lifespan.

Gamification can include transforming experiments into bespoke games, embedding experiments in existing games and extracting data from ongoing ones. The popular game Wordle — in a sense, the world's largest psycholinguistics experiment — has already inspired investigations on topics such as optimization problems in active learning[6] and the contexts in which people cheat[7].



Reproducibility: expect less of the scientific paper

Participants in conventional lab-based studies of human behaviour are often few in number and WEIRD (that is, from Western, educated, industrialized, rich and democratic societies). This leads to results that are statistically imprecise or irreproducible, or that cannot be generalized to other groups.

The massive data sets enabled by gamified science can help to address questions about reproducibility and generalizability. For example, small studies had shown that a person's experience of speaking a tonal language — one that uses pitch, or tones, to distinguish between words, as in Mandarin — alters their ability to perceive musical pitch. However, such work had been conducted largely in Mandarin or Cantonese. To explore tonal languages that are less commonly studied, Jingxuan Liu, now a graduate student at Columbia University in New York City, and one of us (S.A.M.) used data from a popular web-based quiz, 'Test your Musical IQ', to replicate these findings in half a million speakers of such languages, including Ewe, spoken in West Africa, and Burmese, used in Myanmar[8].

Issues of reproducibility and generalizability are particularly acute for scientists who work with hard-to-reach study populations, such as children. Gamified experiments have the potential to encourage participation from children in settings such as schools or museums, rather than requiring a special trip to a lab, which only some families have the time for or interest in

doing. For example, one of us (B.L.) installed a ‘Let’s Draw!’ kiosk at the Children’s Discovery Museum of San Jose in California. Children visiting the museum were asked to draw various things — such as a watch or a tiger — and then play games trying to recognize each other’s drawings. Over the next 18 months, the kiosk collected more than 37,000 drawings from some 8,000 children aged between 2 and 10, creating the world’s largest corpus of children’s drawings and showing how object recognition changes with age[9].



A kiosk at the Children’s Discovery Museum of San Jose, California, collected drawings from thousands of two- to ten-year-olds to reveal how drawing changes systematically with age.Credit: Bria Long

Games played in the lab can also improve the robustness of data. Two of us (A.B.-L. and D.G.W.), working with psycholinguist Joe Toscano at Villanova University in Pennsylvania, wanted to study how people communicate through prosody — the patterns of intonation, pausing and rhythm that convey what information is important, as well as sarcasm and humour. The stilted conversations that participants usually produced in the

lab limited these acoustic cues, and previous results on prosody were inconsistent. We designed puzzles in the game Minecraft that players had to solve by communicating while absorbed with playing. This produced more-natural dialogue and higher-quality data, showing that participants used speech intonation and duration to convey subtle information, such as whether a word was new to the conversation[10].



Invest 5% of research funds in ensuring data are reusable

Compared with lab-based research, online gamified experiments make it easier and cheaper to recruit diverse participants, including those from groups that are under-represented in scientific research. In a study published last year, a gamified experiment on infant-directed speech and song recruited hundreds of people who self-identified across three gender categories and a wide range of household incomes and ethnicities in the United States[11]. More broadly, translating a gamified experiment and distributing it globally on the Internet can help to reduce the deeply English-centric nature of behavioural-science research[12].

The barriers to gamification are receding. Coding an experiment to run online used to require a lab to maintain and develop custom code. But the past few years have seen an explosion in free, open-source tools in mainstream programming languages, such as JavaScript and Python. One of the most common tools for basic human experiments on the Internet, the

JavaScript library jsPsych[13], was launched in 2012 by Josh de Leeuw, a cognitive scientist at Vassar College in Poughkeepsie, New York. It has been used in nearly 1,000 published papers, according to Google Scholar.

The jsPsych tool provides a means for researchers to show their participants stimuli (text, images, audio, video); to ask questions about them; to collect responses in various ways (multiple choice, free text, clicking or tapping an image); and to output data in a structured format. Most importantly, de Leeuw maintains an active user forum and GitHub repository where researchers can discuss questions and collaborate on its codebase. This and similar tools — such as lab.js, Open Sesame and psychTestR — make it straightforward for researchers to start creating behavioural experiments online (see ‘Get into gamification’).

Get into gamification

Using simple tools and making small tweaks to a gamified experiment’s design can vastly increase engagement and expand the number of participants.

• **Take advantage of front-end tools.** Software including jsPsych, psychTestR, lab.js and OpenSesame provides an entry point to web development and gamification. Common frameworks for data collection — such as displaying a series of images and measuring the time taken to press a button — are available by default through plug-ins. A researcher gamifying a lab-based experiment needs only this, a computer and an idea.

• **Think about back-end tools.** Larger-scale experiments require cloud infrastructure that can distribute them, handle data collection and scale dynamically, so that the whole thing doesn’t fall apart if your experiment goes viral. Customizable options such as Pushkin are available, although they currently require more technical expertise than do front-end tools.

• **Provide an incentive.** This could be points, scores or percentile ranks that allow participants to see how well they did, or to learn something about themselves. Pay participants a fair wage for long or onerous tasks.

• **Keep it short.** Try to restrict online games to less than 10 minutes. Some gamified experiments are even quicker.

• **Help it snowball on social media.** Make it easy for participants to share their results on social media — especially as an image or graph. This attracts more participants.

• **Use a hook.** Try adding a storyline or other form of visible progress to keep participants engaged. In one experiment on auditory perception, participants win points as they complete listening tasks to help a bird avatar stay aloft. Nearly 200,000 people have participated so far.

• **Visuals matter.** Putting a little effort into choosing free graphics or paying to design visuals can help to engage participants.

• **Run a pilot.** Before going public with your experiment, try it out on friends and collaborators. They'll be able to tell you what works and what could be improved.

• **Plan and test.** Ensure you have a data-management plan for handling large influxes of data (such as MongoDB or Google Cloud Firestore) if your game goes viral.

• **Collaborate and share.** Developing a project with a gamification lab can benefit everyone. If you develop new functionality, add it to collaborative codebases, as happens with jsPsych plug-ins.

Ups and downs

Gamified experiments have clear weaknesses. Many scientists are used to having total control over their lab environments: they can observe participants' behaviour directly during experiments and check that people are who they say. Critics might be wary of losing this control, or might worry that people will not engage fully with the tests or will warp results by faking their identities, completing games multiple times or participating maliciously using Internet bots.

These criticisms can be partly assuaged. Gamified studies have the potential to engage participants better than lab-based ones can, because they are intrinsically motivating. Indeed, many psychologists have observed all manner of disengagement with lab experiments, such as participants looking at social media or even taking a quick nap, because tasks are often boring and low paid. The fact that many gamified experiments have gone viral on YouTube, Reddit and other social-media platforms provides a proof of concept that participants can become deeply engaged in research — perhaps more so than with conventional approaches, although we need more data to be sure.

Creative approaches to experimental design and data analysis can also give researchers more control over gamified science than might be expected. For example, in a music study aimed at children, cognitive scientist Courtney Hilton at the University of Auckland, New Zealand, and his colleagues (including S.A.M.) studied when participants played a game and compared the distribution of local times with the times adults played. The participants were less likely to play at night, implying that they were indeed children, who are more likely to be asleep at those times[14]. Concerns that participants could be bots can be mitigated by comparing the timing of spikes in recruitment to known events, such as a social-media post promoting a game, and by the use of manipulation checks, which are designed to be easy for a participant to answer, but hard for a scammer or bot. Ultimately, however, some research questions are better suited to the lab. The less-controlled environments used in web-based gamified science can and should complement those of lab work.



The game Wordle has enabled investigation of the contexts in which people cheat.Credit: Claire Welsh/Nature

Another justified critique is that gamified science, when delivered online, leaves out people who lack smartphones, computers and Internet access. But it still offers an improvement on the status quo. Under-resourced communities are already at risk of being left out of lab experiments because labs are concentrated in a small number of places, and participants must spend time and money to visit them. By contrast, smartphones are becoming ubiquitous worldwide. In the United States, three-quarters of adults with household incomes below US$30,000 a year have a smartphone, according to a 2021 survey by the Pew Research Center in Washington DC. The fact that people from lower-income communities are more likely to connect to the Internet using smartphones than with desktop computers means it is a high priority for researchers to develop mobile-friendly experiments that work when Internet connections are slow or patchy.

Gamified experiments, especially large online ones, must also address issues concerning copyright, privacy, data storage and data accessibility. Data-protection and privacy laws differ across countries, and some research studies collect information (such as that on gender or income) that falls under protected categories. To preserve privacy and avoid legal issues, researchers can avoid collecting person-related data, such as IP addresses or other identifying information.

If data are collected by a third-party company in an existing game, the information can be proprietary. A data-use agreement is needed for research and to avoid conflicts of interest. Many universities have legal and ethical teams that can advise on gamified experiments, just as they do for lab ones. For example, when gamifying an experiment involving published music recordings, one of us (S.A.M.) received helpful guidance about US copyright law from Harvard lawyers[15].

Four future steps

For gamification to reach its full potential, four steps need to happen. First, researchers already in the community need to coordinate their development activities to widen access to technology and share technical skills. Labs should make their software publicly available, so that scientists starting out don't need to reinvent the wheel. Although open-science practices are becoming more widespread, thanks to platforms such as GitHub, Zenodo and the Open Science Framework, they are not yet ubiquitous. The community would benefit from pooling resources to create centralized tools and best practices for technology development, as well as running conferences and workshops aimed at increasing access to gamification and citizen-science expertise.



Everyone should decide how their digital data are used — not just tech companies

Second, existing tools should be improved. Many researchers use commercial software such as Qualtrics or SurveyMonkey to conduct online surveys, but no analogous software is freely available for gamified experiments. Front-end tools such as jsPsych are generally not yet optimized for mobile-friendly experiments. Back-end tools that are accessible to scientists without substantial technical expertise are uncommon — but these would jump-start scientists' ability to produce massive-scale gamified experiments, which are still relatively rare. The scientific community should work to develop these resources, and funders worldwide should make it a priority to support these efforts. The US National Science Foundation has already called for digital infrastructure projects as part of efforts to advance reproducibility (see go.nature.com/3xi6bae).

Third, scientists need to look to professions outside research to broaden their technical toolkits. Gamified experiments that produce lots of data could benefit from the use of software and practices that are standard in the technology industry but still underused in science. These include version control using Git (a way to systematically track different versions of code) and storing data in relational databases (such as those using SQL, a language for database management that is ubiquitous in tech companies). Research labs should consider consulting or recruiting artists, graphic designers,

educators and others from diverse fields who can increase the robustness of experiments by improving game design. And hiring or contracting dedicated software developers ensures that technical skills are not lost when graduate students and postdocs leave a lab group. Funders should support these essential positions.

Finally, the community needs to improve the way it dispels inaccurate impressions about gamification, which by its nature risks sounding like a diversion rather than a serious research tool. We think the best way to do so is through science itself. If gamified research continues to enable substantive scientific discoveries, the results will speak for themselves. Scientists should try out gamification on their most creative, pressing and exciting research questions.

The onus is on the community to ensure that these experiments are credible and valuable. Researchers can do so by designing studies that attract deep engagement from the public, by using robust data-management practices, by adopting analytical strategies that demonstrate the validity of gamified data and by sharing data for colleagues to scrutinize. Editors and reviewers can support these efforts by encouraging the publication of high-quality experiments.

In 1986, the sociologist Lee Sproull suggested that researchers consider "a new tool for data collection — electronic mail". Decades after her work on what might be the earliest web-based study[16], collecting data on the Internet is powerful and routine. Gamified science has not yet reached such widespread acceptance, but we think it can, should and will.

Nature **613**, 433-436 (2023)

doi: https://doi.org/10.1038/d41586-023-00065-6

References

1. Hartshorne, J. K., Tenenbaum, J. B. & Pinker, S. *Cognition* **177**, 263–277 (2018).

2. Coutrot, A. *et al. Nature* **604**, 104–110 (2022).

3. Levari, D. E., Gilbert, D. T. & Wilson, T. D. *Psychol. Sci.* **33**, 685–698 (2022).

4. Awad, E. *et al. Nature* **563**, 59–64 (2018).

5. Germine, L. *et al. Psychon. Bull. Rev.* **19**, 847–857 (2012).

6. Brown, K. A. *Npj Comput. Mater.* **8**, 97 (2022).

7. Wormley, A. S. & Cohen, A. B. *Perspect. Psychol. Sci*. https://doi.org/10.1177/1745691622111375 (2022).

8. Liu, J., Hilton, C. B., Bergelson, E. & Mehr, S. A. Preprint at bioRxiv https://doi.org/10.1101/2021.10.18.464888 (2021).

9. Long, B., Fan, J. E., Huey, H. Chai, Z. & Frank, M. C. Preprint at PsyArXiv https://doi.org/10.31234/osf.io/5yv7x (2021).

10. Buxó-Lugo, A., Toscano, J. C. & Watson, D. G. *Discourse Process.* **55**, 305–323 (2018).

11. Hilton, C. B. *et al. Nature Hum. Behav.* **6**, 1545–1556 (2022).

12. Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D. & Majid, A. *Trends Cogn. Sci.* **26**, 1153–1170 (2022).

13. de Leeuw, J. R. *Behav. Res. Methods* **47**, 1–12 (2015).

14. Hilton, C. B., Crowley-de Thierry, L., Yan, R., Martin, A. & Mehr, S. A. *J. Exp. Psychol. Gen*. https://doi.org/10.1037/xge0001289 (2022).

15. Mehr, S. A. *et al. Science* **366**, eaax0868 (2019).

16. Sproull, L. S. *Acad. Mgmt J.* **29**, 159–169 (1986).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00065-6

Work

- **Facing inflation: lab heads tighten supplies budgets** [16 Jan 2023]
  • Group leaders must find ways to efficiently cut costs without sacrificing research quality.
- **Tracking microbes in extreme environments** [16 Jan 2023]
  • Microbial ecologist Gerdhard Jessen studies how climate change might affect microorganisms in deserts, hot springs and hydrothermal vents.

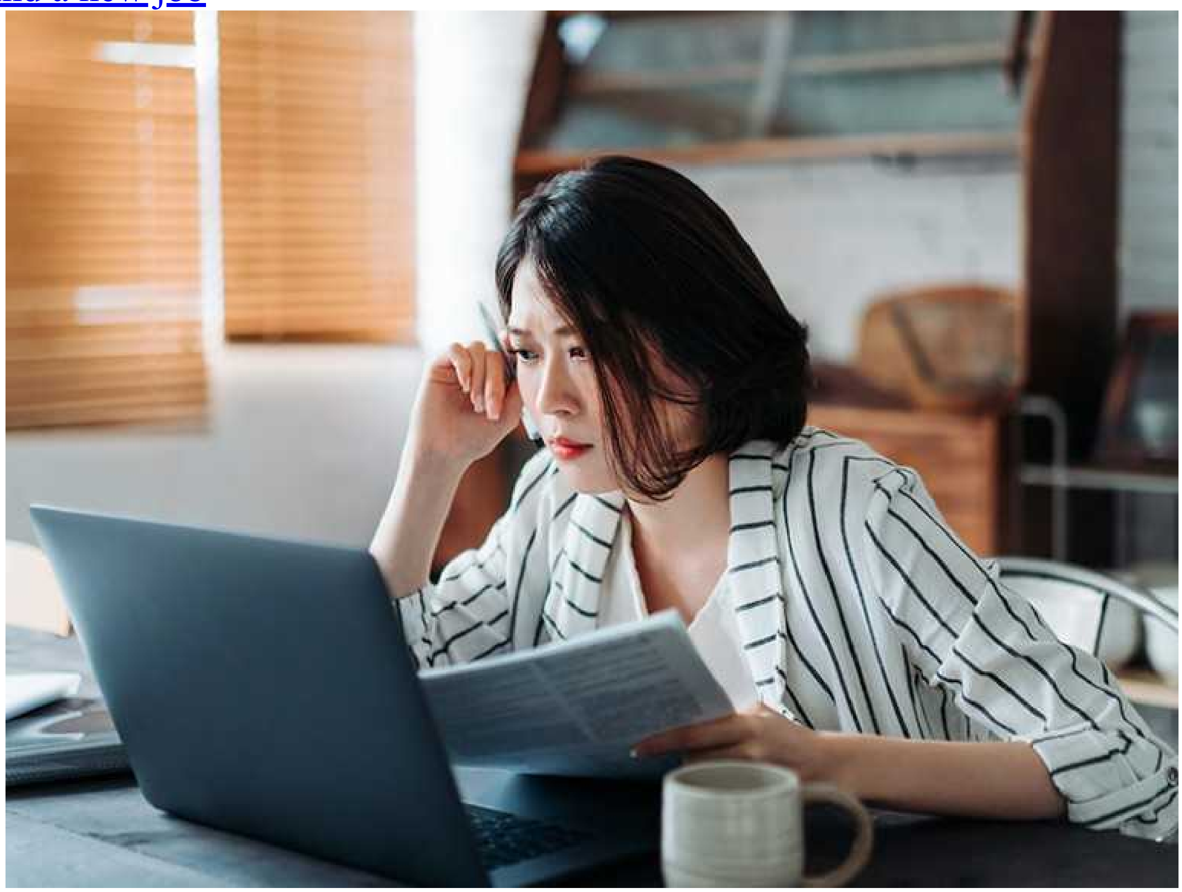
- CAREER FEATURE
- 16 January 2023

Facing inflation: lab heads tighten supplies budgets

Group leaders must find ways to efficiently cut costs without sacrificing research quality.

- Chris Woolston

Find a new job



Research-group leaders must rethink expenses in the face of inflation.Credit: Getty

As the costs of fuel, groceries and housing surge around the world, scientists are fighting inflation at the bench. Reagents, gloves, pipette tips, microscopes and almost every other item needed to conduct science are more expensive than they were just a year ago. And that means that nearly every researcher is feeling the pressure. “Nobody is immune to this economy,” says Tola Olorunnisola, who leads innovation in lab and clinical services at Avantor, an international science-management company in Radnor, Pennsylvania. In late 2022, Olorunnisola visited labs in the Netherlands, Switzerland and Ireland to help researchers facing a cash crunch find ways to stretch their budgets. “Scientists are becoming more conscious of costs,” she says.

Inflation — driven in part by a fractured supply chain that limits the availability of popular items — has become a major issue for Julien Sage, a cancer researcher and geneticist at Stanford University in California. Last October, he and Alyssa Ray, a research administrator at Stanford, ran an audit of the prices of some of the most commonly used items in Sage’s lab. The price of pipette tips has increased by around 10% since January 2021 and by more than 25% since 2018. Also since 2018, prices of syringe filters have increased by 28%, kits for purifying RNA from cells have increased by 46% and medium-sized nitrile gloves (much in demand at the height of the COVID-19 pandemic) have increased by 91%. Altogether, the Sage lab has seen the price of lab supplies increase by 27% since 2018, and the problem is only picking up speed. “We’ve started to suffer from it more in the last 10 months,” Sage says.

High-end equipment is costing significantly more as well. In 2019, the scientific-instrument firm Cole-Parmer in Vernon Hills, Illinois, sold a Qsonica Q700 touch-screen sonicator — a device that uses ultrasonic vibrations to rupture cells — for US$5,520. By January, the same product was offered for more than $7,130. Sage notes that DNA sequencing kits are one of the few items bucking the trend. In early 2022, each kit cost about 3% less than it did in 2018. But that’s hardly enough to offset the increases of other items.

Tough choices

The increase in lab costs has forced scientists to make some difficult choices, says Jim Woodgett, a cancer researcher at the Lunenfeld-Tanenbaum Research Institute in Toronto, Canada. “Scientific budgets are pretty fixed,” he says. “If we pay double for something, it means we’re not buying something else.” Scientists can keep their research projects moving forward, he says, but to avoid overspending on their budgets, they’ll probably need to adjust their buying habits, tweak their experimental designs and take steps to make their labs more efficient.



‘Does anyone have any of these?’: Lab-supply shortages strike amid global pandemic

Woodgett estimates that 20% of his budget is used for lab supplies, with another 60% going to salaries and 20% to mice. Likewise, Sage estimates that lab supplies historically account for roughly 20% of his overall budget, but he says that the balance is shifting. In the past year, Sage says that he has lost one postdoc and one technician position, but that rising costs of supplies helped erase any potential savings. “We still have the same budget,” he says.

Lívia Guadaim, a spokesperson for LatM Life Science, a consulting company headquartered in Lisbon, says that “market dynamics” have posed

a major challenge for scientists. LatM is a sister company of the global lab-supply firm MilliporeSigma in Burlington, Massachusetts. Guadaim declined to provide any specific examples of price increases or to estimate the overall inflation rate for MilliporeSigma products. “We do not disclose details about our pricing policies as we deem them proprietary and competitively sensitive,” she says. She does say that the prices of all products sold by MilliporeSigma are subject to “different levels of impact from things like increasing utility and raw material costs, inflation and logistics”.

Stretched budgets

Inflation has hit almost every part of the scientific enterprise, but funding generally hasn’t kept pace. Sage notes that the amount in dollars of the Outstanding Investigator Award from the US National Cancer Institute (NCI), a division of the National Institutes of Health (NIH), hasn’t changed since the grant was first awarded in 2015. It was $600,000 spread across seven years then, and it’s the same now. “The NCI and the NIH have not paid much attention to inflation,” he says. “That’s clearly going to affect productivity.” But even though prices of lab supplies are rising, researchers are still expected to follow through on the proposed experiments that earned them the grant in the first place, even if those experiments no longer fit within the budget.



UK graduate students demand pay rise from nation’s largest research funder

Woodgett notes that the budgets of funding agencies are generally controlled and limited by federal governments. At a time when research trainees around the world are demanding raises to keep pace with the cost of living, more and more of those funds are going to postdocs and PhD students. Woodgett thinks that funders have not prioritized boosting grant amounts to account for inflation. A post on the NIH website in June 2022 by Michael Lauer, the deputy director for extramural research, noted that the average inflation-adjusted value of research-project grants held relatively stable at around $570,000 from fiscal year 2015 to fiscal year 2021.

Without significant boosts in funding to keep pace with inflation, it’s up to scientists to find creative ways to trim costs. One option, Woodgett says, is to rethink experimental design. “You can evaluate the true need or necessity for a particular reagent or amount of reagent,” he says. “Often, we are quite wasteful. If you downsize your assays by half, you can do double the number of samples.” Some researchers might choose to forgo experiments that seem redundant, but Woodgett says that reviewers of papers tend to ask for those experiments anyway.

Unexpected variables

Woodgett says that researchers can always re-evaluate the reagents and kits that they use for experiments and shop around for cheaper options. But bargain hunting comes with risk, even when it involves rodents. “The bedding can impact the behaviour of the animals,” he says. “If you change the feed a tiny amount and you’re doing metabolic studies, it’s very hard to ensure continuity.”

Sage says that it is often possible to cut 10% from a lab-supply budget simply through smart shopping. “Think about your groceries,” he says. “Saving 10% could be the difference between organic blueberries and normal blueberries. It shouldn’t be that hard.” Yet, like Woodgett, he knows he has to be cautious about any switch. “If the whole lab has been using the same kit, am I going to switch to a kit that’s 10% cheaper? It might introduce a variable that I don’t fully understand.”

Budget-conscious researchers can avoid waste by ensuring that they always have the supplies to finish any experiment they start, Woodgett says. With ongoing supply-chain issues and many common items on back order, it takes planning and foresight to see experiments through to the end. After years of just-in-time ordering, Woodgett had to convert part of his lab to storage, especially for pipette tips and other items that seem to run out at the worst possible times.



How going green can raise cash for your lab

Some distributors might offer discounts for bulk orders, which means scientists can save money by joining forces, Woodgett says. “You can get together with other scientists doing similar work in your city, not just your institution, and arrange a contract,” he says. For example, the US lab-supplies firm Fisher Scientific sells a pack of 20 sterile syringes for $21.75, but a case of 12 packs costs $194, a discount of more than 25%. Large orders are also more likely to be prioritized by the lab-supply company, he adds, which increases the chances that the delivery will be made on time. As a bonus, he says, sharing large shipments of reagents will increase consistency between labs and make it easier to compare and replicate results.

Energy is a significant source of lab costs, especially in Europe and even more so in the United Kingdom, where consumer power prices soared by an average of nearly 90% between October 2021 and October 2022. When

Olorunnisola works with labs that have to pay their own electric bills, she pays particular attention to anything that draws power, including freezers, fume hoods and mass spectrometers. She says it's important to ensure that every instrument is properly maintained, calibrated and used efficiently. To make sure that scientists are getting the most out of their freezers, for example, Olorunnisola often installs sensors that can track how often the door is opened. "You can find out if a freezer hasn't been opened in six months," she says. In such cases, samples can often be consolidated into other freezers or moved to off-site facilities.

Cost-cutting

Scientists keep a careful eye on experimental protocols, but can be less vigilant of waste and inefficiencies, Olorunnisola says. "A lot of scientists that I talk to say they don't know where the redundancies are," she says. A 2020 survey conducted by the publication *Chemical and Engineering News* on behalf of MilliporeSigma of nearly 1,000 researchers found room for improvement (see go.nature.com/3w5syi1).

Of the respondents, 22% said they still used handwritten records to keep track of chemical inventories. More than one-third (35%) of the lab workers spent at least 10% of their time each month searching for samples or reagents. Nearly one in five said that they lose at least 10% of their inventory each month to spoilage and expirations. But standards can change in hard times. "A lot of equipment and reagents have 'sell by' dates that are ridiculous," Woodgett says. "You can stretch those."

A 2022 analysis of antibody concentrates that have nominal expiration dates of one or two years after purchase found that antibodies are still usable an average of six years after the supposed expiration date. The report concluded that "selecting antibodies to be discarded should be based on performance rather than expiration date alone" (A. F. Henwood *Biotech. Histochem.* https://doi.org/jr6h; 2022).

It will probably take more than discounts from lab-supply companies to truly protect scientists from the impact of rising prices, Sage says. Unless something is done on a large scale to either stabilize costs or increase

funding, science is likely to suffer. “The major issue is the downward spiral,” he says. “If you have less money, you’re going to have fewer people or be less productive, which means you’re going to have fewer grants, which means you’re going to have fewer people. That’s probably happening to a lot of labs these days, and the question is: when is it going to stop?”

Nature **613**, 601-602 (2023)

doi: https://doi.org/10.1038/d41586-023-00088-z

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00088-z

- WHERE I WORK
- 16 January 2023

Tracking microbes in extreme environments

Microbial ecologist Gerdhard Jessen studies how climate change might affect microorganisms in deserts, hot springs and hydrothermal vents.

- Virginia Gewin



Gerdhard Jessen is a microbial ecologist at the Austral University of Chile in Valdivia. Credit: Jacopo Pasotti for *Nature*

In 2008, I was investigating the methane bubbling up on the beaches and in shallow waters of Mocha Island, off the coast of central Chile. I became

intrigued by how microorganisms could thrive in methane-rich areas and changed my research focus from marine biology to extreme environments. I wanted to understand how methane acts as a source of energy and carbon for microbes.

Since then, I have explored a number of bizarre environments. In 2010, I went in a submarine down to 200 metres in the Black Sea, one of the world's largest anoxic water bodies. There, I found mats of filamentous bacteria that survive on sulfur compounds.

In 2017, I studied the microbes in Canada's tailing ponds, artificial lakes of water, sand and clay waste that are left behind after petroleum extraction. And I sampled the microorganisms living in 100 °C Antarctic hot springs in 2022.

I came home to Chile in 2018 and began collaborating with an international team researching the geomicrobiology of thermal features, including hot springs, geysers and volcanoes. After travelling with the group to Argentina's active volcanic region, I got funding to explore the microbial communities that exist beneath hydrothermal vents in southern Chile, where the oceanic crust is subducting beneath the continental plate.

In this image, I am in the Atacama Desert in South America, the driest non-polar desert on the planet. I am measuring 80–100 °C steam released from a fumarole containing yellow sulfur, which crystallizes at its opening as the vapour cools. I also sampled sub-surface microbes that are flushed out with the fluids. We'll sequence their DNA to assess the microbial communities and their biological interactions.

My goal is to learn more about subsurface microbes in extreme environments. I want to understand how microbial forces shaped the planet and how these communities might shift in the future with climate change.

Nature **613**, 604 (2023)

doi: https://doi.org/10.1038/d41586-023-00089-y

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00089-y

Research

- **Magnetically mediated hole pairing in fermionic ladders of ultracold atoms** [18 Jan 2023]
  • The direct observation of hole pairing in a doped Hubbard model is demonstrated using ultracold atoms in a quantum gas microscope setting by engineering mixed-dimensional fermionic ladders.
- **Octupole-driven magnetoresistance in an antiferromagnetic tunnel junction** [18 Jan 2023]
  • The authors report observation of tunnelling magnetoresistance in an all-antiferromagnetic tunnel junction consisting of Mn3Sn/MgO/Mn3Sn, laying the foundation for the development of ultrafast and efficient spintronic devices using antiferromagnets.
- **Vertical organic electrochemical transistors for complementary circuits** [18 Jan 2023]
  • Vertical organic electrochemical transistors demonstrating unprecedented performances in both p- and n-type operation modes have been synthesized from new electro-active and ion-permeable semiconducting polymers by the interface engineering of electro-active blend layers.
- **Modern temperatures in central–north Greenland warmest in past millennium** [18 Jan 2023]
  • A reconstruction of temperatures in central and north Greenland from ad 1000 to 2011 shows that that the final decade of this period was on average 1.5 ± 0.4 °C warmer compared to pre-industrial temperatures, accompanied by increased meltwater run-off.
- **FinnGen provides genetic insights from a well-phenotyped isolated population** [18 Jan 2023]
  • Genome-wide association studies of individuals from an isolated population (data from the Finnish biobank study FinnGen) and consequent meta-analyses facilitate the identification of previously unknown coding variant associations for both rare and common diseases.
- **Mono- and biallelic variant effects on disease at biobank scale** [18 Jan 2023]
  • An analysis of biobank data from the FinnGen project examines dosage effects of genetic variants on disease, andidentifies a benefit when considering more complex inheritance in the genetics of common as well as Mendelian diseases.

- **[Financial incentives for vaccination do not have negative unintended consequences]** [11 Jan 2023]
  • Findings from large-scale studies in Sweden and the USA indicate that providing financial incentives for vaccination and informing about state incentive programmes do not have any negative unintended consequences.
- **[Synaptic gradients transform object location to action]** [04 Jan 2023]
  • Analysis of behaviour, physiology, anatomy and connectomics in Drosophila shows how direction-specific visual information is transformed onto downstream premotor networks and converted into appropriate motor responses.
- **[Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir]** [09 Nov 2022]
  • Nirmatrelvir, an oral antiviral targeting the 3CL protease of SARS-CoV-2, has been demonstrated to be clinically useful against COVID-19, but viral resistance to the drug was found to arise readily via multiple pathways in vitro.
- **[A quantitative map of nuclear pore assembly reveals two distinct mechanisms]** [04 Jan 2023]
  • Single-molecule calibrated live microscopy and computational modelling have revealed that human nuclear pore complex assembly takes different pathways during the exit from mitosis and during nuclear growth in interphase.
- **[RNA targeting unleashes indiscriminate nuclease activity of CRISPR–Cas12a2]** [04 Jan 2023]
  • Structural analysis of Cas12a2, a CRISPR-associated nuclease that nonspecifically cleaves ssRNA, ssDNA and dsDNA, reveals a complete activation pathway involved in the abortive infection system protecting cells against invasion.
- **[Cas12a2 elicits abortive infection through RNA-triggered destruction of dsDNA]** [04 Jan 2023]
  • RNA targeting by the Sulfuricurvum type V single-effector nuclease SuCas12a2 drives abortive infection through non-specific cleavage of double-stranded DNA—after recognition of an RNA target through an activating protospacer-flanking sequence, SuCas12a2 efficiently degrades ssRNA, ssDNA and dsDNA.

- Article
- Open Access
- Published: 18 January 2023

Magnetically mediated hole pairing in fermionic ladders of ultracold atoms

- Sarah Hirthe ,
- Thomas Chalopin,
- Dominik Bourgund,
- Petar Bojović,
- Annabelle Bohrdt,
- Eugene Demler,
- Fabian Grusdt,
- Immanuel Bloch &
- …
- Timon A. Hilker

Show authors

Nature **volume 613**, pages 463–467 (2023)

- 2035 Accesses

- 73 Altmetric

- Metrics details

Abstract

Conventional superconductivity emerges from pairing of charge carriers—electrons or holes—mediated by phonons[1]. In many unconventional superconductors, the pairing mechanism is conjectured to be mediated by magnetic correlations[2], as captured by models of mobile charges in doped antiferromagnets[3]. However, a precise understanding of the underlying mechanism in real materials is still lacking and has been driving experimental and theoretical research for the past 40 years. Early theoretical studies predicted magnetic-mediated pairing of dopants in ladder systems[4,5,6,7,8], in which idealized theoretical toy models explained how pairing can emerge despite repulsive interactions[9]. Here we experimentally observe this long-standing theoretical prediction, reporting hole pairing due to magnetic correlations in a quantum gas of ultracold atoms. By engineering doped antiferromagnetic ladders with mixed-dimensional couplings[10], we suppress Pauli blocking of holes at short length scales. This results in a marked increase in binding energy and decrease in pair size, enabling us to observe pairs of holes predominantly occupying the same rung of the ladder. We find a hole–hole binding energy of the order of the superexchange energy and, upon increased doping, we observe spatial structures in the pair distribution, indicating repulsion between bound hole pairs. By engineering a configuration in which binding is strongly enhanced, we delineate a strategy to increase the critical temperature for superconductivity.

Main

Unconventional superconductivity in materials such as heavy fermion systems[11], iron pnictides[12], layered organic materials[13], cuprate superconductors[14] and twisted bilayer graphene[15], arises in the vicinity of magnetically ordered states. A common mechanism consisting of dopant pairing mediated by magnetic fluctuations is thus believed to be at the heart of these superconducting states[2], but a detailed understanding of the underlying physics remains a key problem in quantum many-body physics.

A promising tool to explore strongly correlated quantum systems is analogue quantum simulation using ultracold atoms[16]. Recent experimental progress investigating doped antiferromagnets has been made using single-site

resolved fermionic quantum gas microscopes[17,18,19,20,21,22], which provide microscopic real-space correlations complementary to the spectroscopic and transport measurements performed in solids. They typically simulate the Fermi–Hubbard model, consisting of itinerant spin-1/2 fermions within a single band of a periodic lattice. Even though the two-dimensional Fermi–Hubbard model displays many characteristics also found in the high-temperature (high-T_C) superconducting cuprates[3], the existence of pairing and superconductivity in this model remains a subject of debate[23,24].

To shed light on the pairing mechanism in doped Mott insulators, several theoretical studies considered doped Fermi–Hubbard and t−J ladders (see Eq. (1))[4,5,6,8,25], in which accurate numerical solutions can be obtained using the density-matrix renormalization group (DMRG) method[26]. In solid-state experiments, ladder materials have also been shown to display superconductivity[27,28,29]. A paradigmatic case for theoretical investigation, which exhibits large binding, is the regime in which interchain magnetic exchange is larger than single-particle interchain hopping[7]. These parameters could, however, not be justified microscopically for condensed matter systems and are unphysical in the framework of a pure Fermi–Hubbard system (Methods). The key motivation for our work was to provide an experimental realization of this system that has been considered only a theoretical abstraction. We achieve this by extending Fermi–Hubbard ladders at large interactions with a potential offset, effectively realizing a mixed-dimensional (mixD) system[10].

The essential physics of our experiment is captured (Methods) by the t−J Hamiltonian

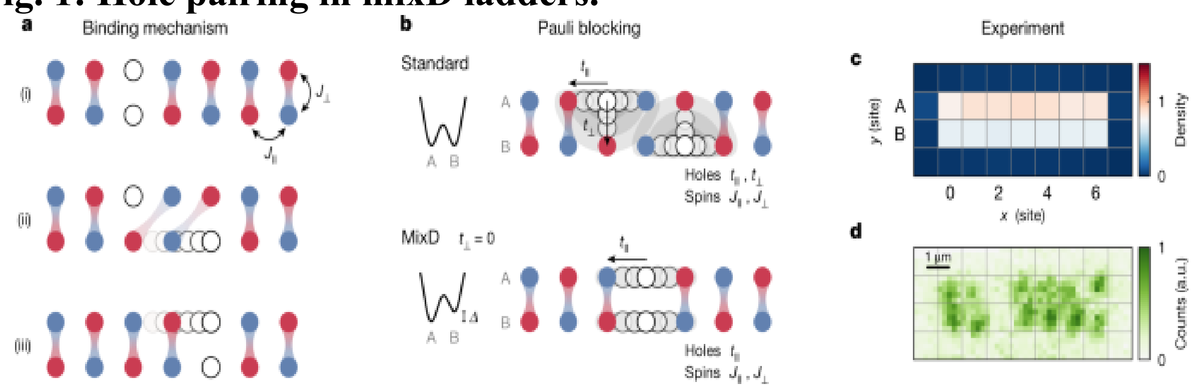
$$\hat{{\mathcal{H}}}=\sum _{\langle i\,,j\rangle ,\,\sigma }\hat{{\mathcal{P}}}(-{t}_{ij}\,{\hat{c}}_{i,\sigma }^{\dagger }{\hat{c}}_{j,\sigma }+{\rm{h.}}{\rm{c.}})\hat{{\mathcal{P}}}+\sum _{\langle i\,,j\rangle }{J}_{ij}\left({\hat{{\bf{S}}}}_{i}\cdot {\hat{{\bf{S}}}}_{j}-\frac{{\hat{n}}_{i}{\hat{n}}_{j}}{4}\right),$$
(1)

where $\hat{{\mathcal{P}}}$ projects to the subspace without double occupancies; the hopping energy is $t_{ij} = t_{\parallel}$ ($t_{\perp}$) and the superexchange energy

is $J_{ij} = J_{\parallel}$ ($J_{\perp}$) for nearest-neighbour sites i, j on the same leg (rung). ${\hat{c}}_{i,\sigma }^{\dagger }$ (${\hat{c}}_{i,\sigma }$) denotes the creation (annihilation) operator of a fermion on site i with spin $\sigma = \uparrow,\downarrow$; ${\hat{{\bf{S}}}}_{i}$ and ${\hat{n}}_{i}$ are the on-site spin and density operators and h.c. means Hermitian conjugate. The mixD system is described by equation (1) for $t_{\perp} = 0$, which we realize by suppressing tunnelling, but not spin exchange, using a potential offset along the rungs. We also realize the bare Fermi–Hubbard system without additional potential offset and $t_{\perp} > J_{\perp}$ (in the following this is denoted as a standard ladder).

The pairing of holes in a model exhibiting only repulsive interactions[30,31] can be understood from the competition of hole delocalization and spin order along the rungs, as illustrated in Fig. 1a. If a hole moves through the system, it displaces these spins creating an energetically unfavourable magnetic configuration. Therefore, a single hole becomes dressed by a cloud of disturbed correlations and forms a polaron with reduced mobility[19,20]. If a second hole moves along with the first hole, it can restore the order in the spin sector, causing the two holes to form a highly mobile bound pair.

Fig. 1: Hole pairing in mixD ladders.



a, Binding mechanism in the t–J ladders. Depicted are ladder systems with spin exchange $J_{\perp} \gg J_{\parallel}$ that form strong singlet bonds along the rungs. When a single hole from (i) moves through the system, as illustrated in (ii), it breaks the spin order by displacing the singlet bonds. (iii) The magnetic energy cost can be avoided if the second hole restores the spin order by moving together with the first hole. **b**, Pauli blocking of holes. Owing to their fermionic nature, holes repel each other along all directions according to the tunnelling amplitudes $t_{\perp}$ and $t_{\parallel}$. Close-distance hole pairs are thus

energetically unfavourable. In mixD systems, a potential offset $\varDelta$ between the two legs suppresses tunnelling $t_\perp$ and Pauli repulsion only occurs along the legs. Holes on the same rung can thus benefit from the binding mechanism, forming tightly bound pairs with a large binding energy. **c**, Average density of the mixD $L = 7$ ladder system with $\varDelta \approx U/2$. **d**, Single experimental shot with two holes on the same rung, exemplifying the bunching of holes in the mixD system. a.u., arbitrary units.

Source data

However, this process only dominates in the mixD case, whereas in the standard ladders $t_\perp$ is the dominating energy scale. In the latter, binding competes with the repulsion of holes due to Pauli blocking, rendering tight pairs energetically unfavourable (Fig. 1b). Pairing between holes can still occur, but only with a small binding energy $E_b \ll J_\perp$[32,33] and at large pair sizes[34]. By realizing mixD ladders, we strongly suppress the Pauli repulsion between holes along the rung (Fig. 1b), thus engineering a system with strong hole attraction. This binding mechanism is protected by the spin gap and thus persists up to high temperatures of the order $J_\perp$.

We realize ladders of length $L = 7$ in our Fermi-gas microscope with independently tunable optical lattices and single-site resolved optical potential shaping. The mixD t–J system is derived as an effective model from fermionic atoms in a two-leg ladder-shaped lattice potential described by the Hubbard parameters U (on-site interaction), $({\widetilde{t}}_{\parallel },{\widetilde{t}}_{\perp })$ (tunnelling amplitudes) and with a potential offset $\varDelta$ between the two legs. For large $(U/{\widetilde{t}}_{\parallel },U/{\widetilde{t}}_{\perp })$, the system is effectively described by the t–J model and its distinct parameters as in equation (1). For $U > \varDelta > {\widetilde{t}}_{\perp }$, tunnelling along the rungs is suppressed to $t_\perp = 0$, whereas tunnelling along the leg is unaffected, ${t}_{\parallel }={\widetilde{t}}_{\parallel }$, and spin exchange is given by ${J}_{\perp }=2{\widetilde{t}}_{\perp }^{2}/(U+\varDelta)+2{\widetilde{t}}_{\perp }^{2}/(U-\varDelta)$(refs. [35,36]) and ${J}_{\parallel }=4{\widetilde{t}}_{\parallel }^{2}/U$.

Our experiment begins by preparing a balanced mixture of the lowest two hyperfine states of ^{6}Li, which we load into an engineered ladder geometry similar to our previous work[37]. We first load atoms into uncoupled legs of equal potential and then apply an optical potential offset Δ to one of the legs using light shaped by a digital micromirror device (DMD). We subsequently connect both legs by slowly lowering the lattice potential in the rung direction (for details see Methods). The offset between the two legs prevents the atoms from tunnelling along the rungs, such that we end up with roughly equally populated legs. Occupation readout is achieved with single-site spin and charge resolution[38]. To highlight the influence of Pauli repulsion on hole pairing, we compare the mixD $\Delta \approx U/2$ case with the standard ladder system at $\Delta = 0$. Both systems are realized in the strong rung-singlet regime, with $J_\perp/J_\parallel = 21(5)$ in the mixD case with enhanced spin exchange and $J_\perp/J_\parallel =$ 16(3) in the standard ladders, where numbers in parentheses denote the uncertainty on the last digit. The mixD system is governed by the energy scales of rung spin exchange and leg tunnelling with $t_\parallel/J_\perp = 0.7(1)$. The strong rung coupling keeps the spatial extent of hole pairs small while staying in the regime in which spin correlations and hole motion compete on a comparable energy scale. For larger $t_\parallel$, the binding energy is even expected to grow[10] but becomes harder to observe in systems of limited sizes because of increased pair sizes.

Figure 1c shows the average density of the mixD system.
Quantum fluctuations, which are biased towards the lower leg, as well as preparation imperfections lead to a small residual density imbalance between the two legs. For the data analysis (except Fig. 1c) we only take into account ladders without double occupancies (doublons), such that holes arising from doublon–hole fluctuations do not contribute to our observations. Unless otherwise mentioned, we furthermore restrict the data to realizations with two to four holes per ladder and limit the occupation imbalance between the legs to one hole.

To probe the pairing of holes in our system, we evaluate the hole–hole correlator

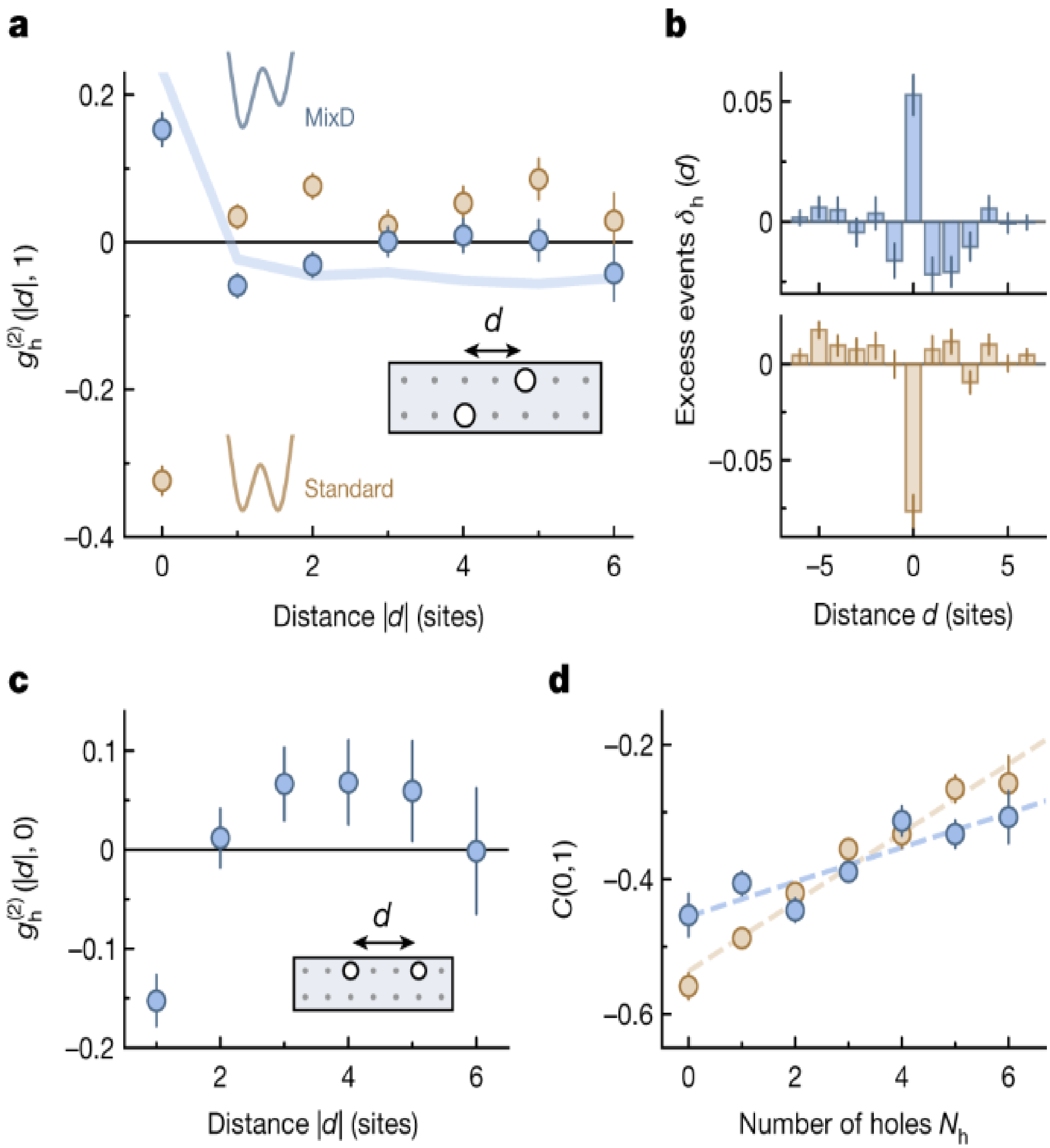
$${g}_{{\rm{h}}}^{(2)}({\boldsymbol{r}})=\frac{1}{{{\mathscr{N}}}_{{\boldsymbol{r}}}}\sum _{{\boldsymbol{i}}-

{\boldsymbol{j}}={\boldsymbol{r}}}\left(\frac{\langle {\hat{n}}_{{\boldsymbol{i}}}^{{\rm{h}}} {\hat{n}}_{{\boldsymbol{j}}}^{{\rm{h}}}\rangle }{\langle {\hat{n}}_{{\boldsymbol{i}}}^{{\rm{h}}}\rangle \langle {\hat{n}}_{{\boldsymbol{j}}}^{{\rm{h}}}\rangle }-1\right),$$

with normalization \({{\mathcal{N}}}_{{\boldsymbol{r}}}\) the number of sites ***i***, ***j*** at distance ***r*** and \({\hat{n}}_{{\boldsymbol{i}}}^{{\rm{h}}}\) the hole-density operator at position ***i*** . The function \({g}_{{\rm{h}}}^{(2)}({\boldsymbol{r}})\) is a connected two-point density correlator that is negative if the presence of a hole at position **i** makes it less likely to find a second hole at distance ***r*** and positive if it makes it more likely. The correlator is bounded by \(-1\le {g}_{{\rm{h}}}^{(2)}({\boldsymbol{r}})\le (1/{n}_{{\rm{h}}}-1)\) with hole density $n_{\rm h} = N_{\rm h}/2L$, where $N_{\rm h}$ is the number of holes in the system.

By evaluating the correlation for holes in opposite legs $\boldsymbol{r} = (d, 1)$, we observe a strong positive signal at distance $|d| = 0$ in the mixD system of \({g}_{{\rm{h}}}^{(2)}(0,1)=0.15(2)\). This corresponds to two holes bunching on the same rung (Fig. 2a). The fast decrease of correlations for $|d| > 0$ indicates that the holes are in a tightly bound state. We find a minimum at $|d| = 1$, which we attribute to the effect of additional holes in the system. These holes are repelled from the hole pair (see also Fig. 4), leading to the weak modulation at larger distances which dominates over the extent of the hole pair at short distances.

Fig. 2: Hole pairing in mixD versus standard ladders.



a, Hole–hole correlator \({g}_{{\rm{h}}}^{(2)}(d,1)\) between sites on opposite legs (as illustrated in the inset) for mixD (blue) and standard (brown) ladders with two to four holes per ladder. The strong correlation at $d = 0$ corresponds to two holes on the same rung. Correlations at this distance are strongly enhanced in the mixD system (pairing) and strongly suppressed in the standard ladders (repulsion). The blue line is calculated using MPS at finite temperature $k_B T = 0.8\, J_{\perp}$ and corrected by the experimental detection fidelity (Methods). **b**, Excess events $\delta_h(d)$ of the

same data, that is, the likelihood of finding holes at distance d compared with the infinite-temperature distribution. **c**, Hole–hole correlation on the same leg $g_{\rm h}^{(2)}(d,0)$ in the mixD system, showing that holes repel each other within the same leg. A finite-size offset correction has been applied to this subfigure (Methods). **d**, Spin–spin correlations $C(0, 1)$ for spins on the same rung depending on the number of holes in the system. The lines represent linear fits. The larger slope indicates that the spin order of the standard system (brown) is more strongly disturbed by holes than the spin order of the mixD system (blue), where paired holes leave the spin order largely unperturbed. The error bars denote one s.e.m. and are smaller than the marker when not visible. Error bars in **a**, **c** and **d** are estimated using bootstrapping.

Source data

By contrast, in the standard ladders of $t_\perp > J_\perp$, a strong repulsion of holes from the same rung is the dominant feature leading to a negative $g_{\rm h}^{(2)}(0,1)$ (Fig. 2a). This shows that tightly bound pairs are energetically unfavourable in the standard ladder system.

The attraction (mixD) and repulsion (standard) of holes are also visible in the occurrences of hole distances. In Fig. 2b we plot the histograms of holes found at a mutual distance d as described by $\delta_{\rm h}(d)=\sum_{\boldsymbol{i}-\boldsymbol{j}=(d,1)}(\langle \hat{n}_{\boldsymbol{i}}^{\rm h}\hat{n}_{\boldsymbol{j}}^{\rm h}\rangle - n_{\rm h}^{2})$, where subtracting the global hole density $n_{\rm h}$ removes the uncorrelated distribution. These excess events $\delta_{\rm h}(d)$ can be interpreted as the likelihood of the hole distance d occurring beyond the probability of a random distribution.

The hole–hole correlator on the same leg (Fig. 2c) shows the effect of hole mobility in the mixD system. It exhibits a minimum at nearest neighbours caused by the Fermi repulsion of holes due to the leg tunnelling being proportional to $t_\parallel$ and a broad maximum around $|d| = 4$. This is the largest mutual distance the two holes can assume without occupying the edge of the

system, which is energetically expensive owing to the hard walls blocking hole movement.

We investigate the magnetic origin of the pairing mechanism using the spin correlator

$$C({\boldsymbol{r}})=\frac{1}{{{\mathscr{N}}}_{{\boldsymbol{r}}}}\sum _{{\boldsymbol{i}}-{\boldsymbol{j}}={\boldsymbol{r}}}4({\langle {\hat{S}}_{{\boldsymbol{i}}}^{z}{\hat{S}}_{{\boldsymbol{j}}}^{z}\rangle }_{{\rm{s}}}-{\langle {\hat{S}}_{{\boldsymbol{i}}}^{z}\rangle }_{{\rm{s}}}{\langle {\hat{S}}_{{\boldsymbol{j}}}^{z}\rangle }_{{\rm{s}}}) \quad (2)$$

where $\langle \,{\rangle }_{{\rm{s}}}$ denotes the expectation value for singly occupied sites. In the doped mixD system we find strong nearest-neighbour spin correlations along the rung of $C(0, 1) = -0.38(1)$, indicating a high singlet fraction, as well as a significant coupling of these bonds along the leg with nearest-neighbour spin correlations of $C(1, 0) = -0.10(1)$. Resolving the rung spin correlations through the number of holes in the system shows a decrease in correlation strength with growing hole number (Fig. 2d). This can be explained by unpaired holes breaking singlet bonds because of their mobility along the ladder. The standard system, which does not display pairing, shows a more rapid loss of correlation strength compared to the mixD system, in which a significant fraction of holes is bound in pairs. This behaviour is directly related to the magnetic origin of pairing, as illustrated in Fig. 1. We attribute the lower correlation strength of the mixD system in the absence of holes to heating caused by the application of the tilt. The much stronger effect of hole doping on spin correlations is, however, unaffected by a small difference in temperature. The slight alternating behaviour of $C(0, 1)$ at low hole numbers in the mixD system, in particular the strong spin correlations for two holes, is reminiscent of the low-temperature behaviour for which rung pairs are dominant over thermal excitations and only odd hole numbers lead to unpaired holes (Supplementary Information).

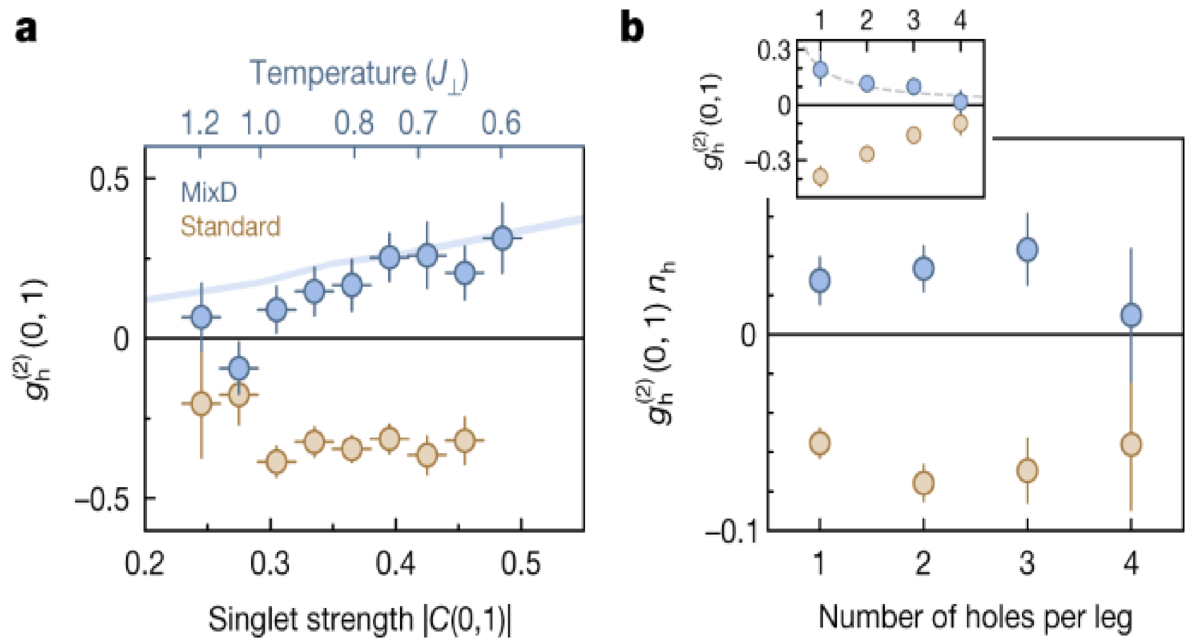
We estimate the binding energy of the paired state in the mixD system from the experimental data, by comparing it with an analytically tractable effective Hamiltonian. The approach is based on the assumption that the system is reasonably close to the uncoupled rung limit and the bound state can thus be described by two holes on the same rung (for details on the calculation, see Methods). Using the measured and detection-fidelity-corrected probability to find a rung pair and our estimated temperature of $k_{\rm B}T = 0.77(2)\,J_\perp$ (Methods), we find an experimental binding energy of

$${E}_{{\rm{b}}}=0.82(6)\,{J}_{\perp }\,.$$

This is consistent with DMRG calculations giving a binding energy of $({E}_{{\rm{b}}}^{{\rm{theo}}}=0.81\,{J}_{\perp })$ and boosts the binding energy by an order of magnitude compared with standard ladders at the same interaction strength and with symmetric coupling (see also Methods). Such a substantial increase in pairing strength is an essential ingredient in the quest for higher temperature superconductivity.

To gain a better understanding of the system dependencies, we probe the influence of temperature and doping on the rung pairing strength. In our experimental regime, nearest-neighbour spin correlations along the rung stand in strong direct relation with the temperature of the system and can therefore be used as an effective thermometer. A mapping between the two is obtained by comparing average nearest-neighbour spin correlations to finite-temperature matrix product state (MPS) calculations (Methods). We observe that the rung hole–hole correlation ${g}_{{\rm{h}}}^{(2)}(0,1)$ increases with increasing spin correlation strength (Fig. 3a), and the onset of pairing occurs in the experimental system around temperatures of the order of the spin exchange $J_\perp$. The highest correlations reached, ${g}_{{\rm{h}}}^{(2)}=0.3(1)$, are still smaller than the theoretically achievable values of ${g}_{{\rm{h}}}^{(2)} > 1.2$ for very low temperatures, showing the temperature limitations of the experiment. The repulsion of the standard system remains mostly constant in our temperature regime, as it is governed by the energy scale $t_\perp \gg J_\perp, k_{\rm B}T$.

Fig. 3: Temperature and doping dependence of hole pairing.



a, Rung hole–hole correlation \({g}_{{\rm{h}}}^{(2)}(0,1)\) for the mixD (blue) and standard (brown) ladders binned by the rung spin correlations *C*(0, 1) of the system. The temperature of the mixD system (top axis) is estimated by comparing the spin correlations (lower axis) with the theoretical values. The solid line is calculated using MPS and is corrected by the experimental detection fidelity. We see unbinding of pairs at low singlet strength, that is, high temperature. **b**, The hole correlator scaled with the hole density \({g}_{{\rm{h}}}^{(2)}(0,1)\,{n}_{{\rm{h}}}\) depending on the number of holes per leg for the mixD (blue) and standard (brown) ladders. Within our error bars, we find the hole binding to be independent of doping. The inset shows the correlator \({g}_{{\rm{h}}}^{(2)}(0,1)\), where the dashed line is a fit with the inherent $1/n_h$ scaling of the correlator. Error bars denote the bin width of the spin correlations (**a**) and the s.e.m. of the correlator (**a** and **b**).

Source data

To explain the effect of doping on the pair binding, we adjust the correlator \({g}_{{\rm{h}}}^{(2)}\) to compensate for the intrinsic density-dependent $1/n_h$ scaling. The resulting correlator \({g}_{{\rm{h}}}^{(2)}{n}_{{\rm{h}}}\) does not significantly change on increasing the number of holes per leg (Fig. 3b). This is in agreement with a system of independent pairs for the mixD system. For the standard system, hole repulsion is

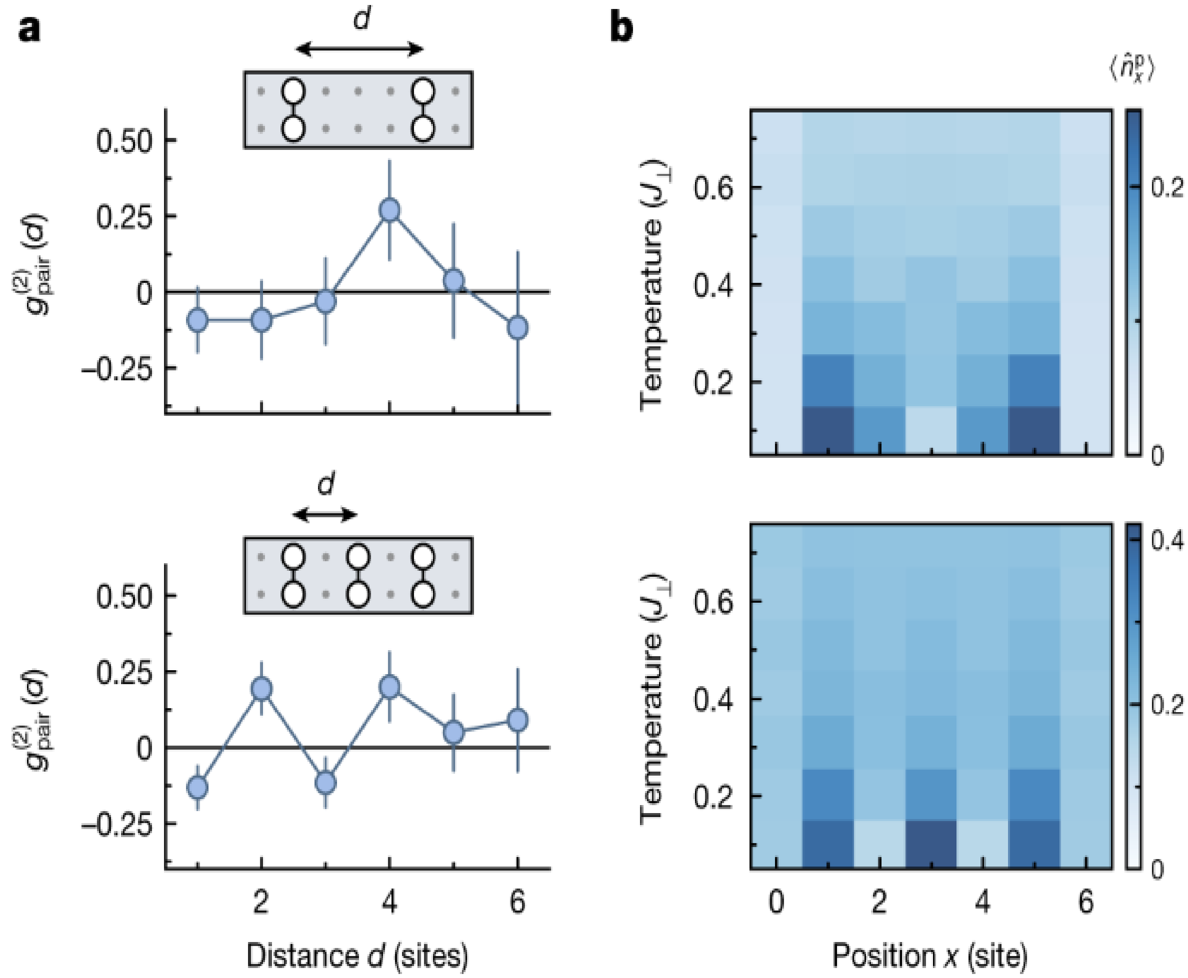
determined by the tunnelling strength $t_\perp$ and is therefore also not significantly influenced by density.

We further investigate the behaviour of several pairs, as their interplay is a key aspect for the competition between superconductivity and charge (density) order[23,39]. For simplicity, we identify the bound state as two holes occupying the same rung. We thus define the pair operator ${\hat{n}}_{x}^{{\rm{p}}}$ which is equal to 1 if both sites of rung x are occupied by a hole, and 0 otherwise. Pair interactions are then quantified by the pair–pair correlator

$${g}_{{\rm{pair}}}^{(2)}(d)=\frac{1}{{{\mathcal{N}}}_{d}}\sum _{x}\left(\frac{\langle {\hat{n}}_{x}^{{\rm{p}}}{\hat{n}}_{x+d}^{{\rm{p}}}\rangle }{\langle {\hat{n}}_{x}^{{\rm{p}}}\rangle \langle {\hat{n}}_{x+d}^{{\rm{p}}}\rangle }-1\right),$$
(3)

in analogy with the hole–hole correlator ${g}_{{\rm{h}}}^{(2)}$. We evaluate ${g}_{{\rm{pair}}}^{(2)}$ on a subset of our data in which at least two pairs are present in the system. In the case of ladders containing four to five holes (Fig. 4a), we find a peak in the correlator at distance $|d| = 4$, whereas it peaks at distances $|d| = 2$ and $|d| = 4$ for the ladders containing six to seven holes. The pairs thus arrange in a spatial structure for which they maximize both their mutual distance and the distance to the edges of the system. This repulsion between pairs is an indication of the comparably high mobility of pairs in mixD settings, which could be further enhanced by increasing the leg tunnelling $t_\parallel$ in larger systems. Such highly mobile pairs can potentially reach very high critical temperatures[10] and can be an important ingredient for high-T_C superconductivity.

Fig. 4: Distribution of rung hole pairs in the mixD system.



a, Measured pair–pair correlation $g_{\rm{pair}}^{(2)}(d)$ of rung hole pairs in the experimental system. The upper plot shows the pair–pair correlation for four to five holes, that is, up to two pairs, in the system. The lower plot shows the pair–pair correlation for six to seven holes, that is, up to three pairs in the system. A finite-size offset correction has been applied to the curves (Methods). Error bars were estimated using bootstrapping. **b**, Theoretical (MPS) results for the density of rung pairs in the system for temperatures from $0.1\,J_\perp$ to $0.7\,J_\perp$. The upper plot shows the pair density for four holes. The lower plot shows the pair density for six holes. In both cases, the pairs maximize their respective distance, while also avoiding the edge of the system.

Source data

To investigate the pair interactions numerically, we perform finite-temperature MPS simulations on a system with the same parameters as in the experiment ($L = 7$, $t_{\parallel}/J_{\perp} = 0.7$). Here pair interaction is directly revealed by the density of pairs $\langle {\hat{n}}_{x}^{{\rm{p}}}\rangle$, which shows strong spatial dependence in the low-temperature regime (Fig. 4b) with an average distance of $|d| = 4$ ($|d| = 2$) for a system with four (six) holes. The presence of sharp edges fixes the phase of the density modulation, thus leading to its visibility in the pair density as a direct consequence of our open boundary conditions and small system size. Such a modulation of the pair distribution is reminiscent of Friedel oscillations of indistinguishable fermions near an impurity[40] and of charge-density waves[41]. Larger systems are needed to distinguish between these effects.

In this work, we have demonstrated the direct observation of hole pairing in a quantum gas microscope setting. We have realized a paradigmatic model that reaches hole binding at high temperatures close to the spin-exchange energy and small pair size by engineering doped, mixD fermionic ladders. We confirm the effective mixD description, predicting that the suppression of Pauli repulsion enables the formation of a bound state.

This allows us to experimentally investigate the pairing mechanism based on hole motion and magnetism, thereby emphasizing the relevance of magnetic correlations as a potential origin of the charge carrier pairing underlying unconventional superconductivity. Furthermore, we have seen signs of significant mobility of the bound pairs through their repulsive interaction. Our approach with strongly bound and mobile pairs can lead to high critical temperatures and therefore can push the current temperature limits of high-T_C superconductors[10]. Possible techniques to modify the interplay of kinetic and exchange energies in solid-state systems in the spirit of the mixD setting include changing lattice geometry and spin polarization[42], or controlling kinetic energy and the topology of electrons by varying the twist angle in moiré systems such as twisted graphene[43]. Our approach might also be realized in real materials as dynamical superconductivity using Floquet engineering to alter the effective exchange interactions[44]. Within the scope of quantum simulations, our technique can be readily extended to higher dimensions using bilayer quantum gas microscopes[38,45,46], and to even higher binding energies and pair mobility in larger systems at $t_{\parallel} > J_{\perp}$, or

stripe formation at higher leg numbers[47]. Our results thus pave the way for the measurement of collective phases of bound pairs such as crystallization and superfluidity, and shed light on their competition[23,39,48,49].

Methods

Experimental sequence

In each experimental run, we prepare a cold atomic cloud of ^{6}Li in a balanced mixture of the lowest two hyperfine states $\left|F=1/2,{m}_{F}=\pm 1/2\right\rangle$. For evaporation, we confine the cloud in a single layer of a staggered optical superlattice along the *z* direction with spacings $a_s = 3$ μm and $a_l = 6$ μm, and depths ${V}_{{\rm{s}}}=51\,{E}_{{\rm{R}}}^{{\rm{s}}}$ and ${V}_{{\rm{l}}}=120\,{E}_{{\rm{R}}}^{{\rm{l}}}$, where ${E}_{{\rm{R}}}^{\alpha }={h}^{2}/(8M{a}_{\alpha }^{2})$ denotes the recoil energy of the respective lattices (α = s, l), and *M* is the mass of an atom. The atoms are harmonically confined in the *xy* plane and the evaporation is performed by ramping up a magnetic gradient along the *y* direction (see ref. [38]). The final atom number is adjusted by the evaporation parameters.

We adiabatically load the cloud into an optical lattice in the *xy* plane with spacings $a_x = 1.18$ μm and $a_y = 1.15$ μm. Simultaneously, we apply a repulsive potential using a DMD, which both compensates for the harmonic confinement of the Gaussian-shaped lattice beams and shapes the system into a geometry of four 2 × 7 ladders following the procedure described in ref. [37]. The DMD is also used to apply a spin-independent optical potential offset Δ between sites along the rung direction. The loading is performed in three steps (Extended Data Fig. 1). (1) A first stage, in which the two legs of each ladder are nearly disconnected, is reached by ramping the lattice depths to ${V}_{x}=3\,{E}_{{\rm{R}}}^{x}$ and ${V}_{y}=20\,{E}_{{\rm{R}}}^{y}$ in 100 ms. (2) The optical potential offset Δ (see Potential offset calibration) is applied to one leg of each ladder by instantaneously (<20 μs) switching the pattern of the DMD. (3) The lattice depths are ramped linearly to their final values ${V}_{x}=12\,$

${E}_{{\rm{R}}}^{x}$ and ${V}_{y}=6\,{E}_{{\rm{R}}}^{y}$ in 100 ms.

Interactions between the atoms are set by the *s*-wave scattering length a_s, which we adjust by applying a magnetic field close to the broad Feshbach resonance of ^{6}Li around 830 G. The scattering length is increased from $a_s \approx 350a_{\rm B}$ during evaporation, with $a_{\rm B}$ being the Bohr radius, to $a_s \approx 1{,}310a_{\rm B}$ in the final configuration. The resulting system is described by the Fermi–Hubbard model and an additional potential offset Δ. Our parameters are the repulsive on-site interaction $U = h \times 4.29(10)$ kHz, tunnelling ${\widetilde{t}}_{\parallel }=h\times 78(10)$ Hz and ${\widetilde{t}}_{\perp }=h\times 303(23)$ Hz and the offset $\Delta \approx 0.5U$ or $\Delta = 0$ depending on the configuration (mixD or standard). As $U/{\widetilde{t}}_{\perp },U/{\widetilde{t}}_{\parallel }\ge 14$, the system can be effectively described by the t−J model (see also 'From the Fermi–Hubbard to the t−J model'). Note that we use ${\widetilde{t}}_{\perp },\,{\widetilde{t}}_{\parallel }$ for the tunnelling parameters in the Hubbard model, and $t_\perp$, $t_\parallel$ in the t−J model. Along the legs, the tunnel coupling is independent of Δ and is ${t}_{\parallel }={\widetilde{t}}_{\parallel }=h\times 78(10)$ Hz, yielding a spin exchange of $J_\parallel = h \times 5.7(1.5)$ Hz. Along the rungs, the mixD system ($\Delta/U \approx 0.5$) yields $t_\perp = 0$ and an enhanced spin exchange $J_\perp = h \times 114(42)$ Hz. Without the potential offset, that is, in the standard system ($\Delta = 0$), tunnelling is unaffected, leading to ${t}_{\perp }={\widetilde{t}}_{\perp }=h\times 303(23)$ Hz and $J_\perp = h \times 86(13)$ Hz.

Potential offset calibration

We realize the mixD system by applying a local spin-independent light shift Δ to one of the legs on each ladder. The amplitude is directly proportional to the light intensity, which is controlled by the DMD. Calibration of the offset is performed by running the experimental sequence described above for different light intensities, and measuring the density of doubly occupied sites (doublons) in the system. An increase of doublons is seen when $\Delta \approx U$, that is, when the lowest band of the upper leg becomes resonant with the interaction band of the lowest leg (Extended Data Fig. 2). This calibration

was repeated several times throughout data collection, with typical shift of the doublon peak of around 10%. We attribute these calibration differences to drifts in the beam shape of the light that is sent to and diffracted from the DMD, yielding an overall estimation of the uncertainty on $\varDelta$ of about ±15%. Such uncertainty in $\varDelta$ is not critical for realizing a mixD setting and mostly influences the value of $J_{\perp}$.

Suppression of rung tunnelling

The potential offset $\varDelta \gg {\widetilde{t}}_{\perp }$ between the two legs shifts the energy levels between neighbouring sites and thus suppresses tunnelling along the rungs[47]. Doublon–hole pairs, however, become biased in the mixD system and appear predominantly as double occupancy on the leg with lower potential and corresponding empty site on the upper leg. Although in the standard system doublon–hole pairs along the rung appear with probability $\propto {({\mathop{t}\limits^{ \sim }}_{\perp }/U)}^{2}$ (ref. [46]), the potential offset in the mixD case lowers the energy difference between the singly occupied state and the doublon–hole pair to $U-\varDelta$. This effect can be seen in the density of the mixD system in Fig. 1c. In Extended Data Fig. 3, we plot the same data after removing ladders containing double occupancies. The density imbalance mostly disappears, indicating minimal tunnelling during preparation.

Detection

The data presented in the main text originate from two types of measurement: (1) charge-resolved and (2) spin-charge-resolved measurements. In both cases, the detection procedure starts by ramping the *xy* lattices to $43\,{E}_{{\rm{R}}}^{xy}$ within 250 µs, effectively freezing the occupation configuration. In the case of spin-resolved measurements (2), a Stern–Gerlach sequence separates the two spin species into two neighbouring planes of the vertical superlattice, which are then separated by 21 µm from one another using the charge pumping technique described in ref. [38]. Finally, fluorescence images are taken using Raman sideband cooling in our dedicated pinning lattice with an imaging time of 1 s (ref. [50]). For a charge-only measurement (1), only one plane is populated by atoms, whereas in the case of a fully resolved measurement (2) two planes

are populated by the two different spin species. The fluorescence light of the atoms is then collected through a high-resolution objective and imaged onto a camera. For a fully spin-resolved measurement (2), the fluorescence of both planes is collected simultaneously and imaged on the camera, allowing the reconstruction of the atomic distribution of both spins with a single exposure. A charge-only measurement only allows the reconstruction of the atomic configuration, without any spin information.

The imaging technique and the pumping procedure both impact our overall detection fidelity. The imaging fidelity, which takes into account atom losses and atom displacement during the imaging procedure, is estimated by comparing two consecutive fluorescence pictures of the same atomic distribution, and we obtain an average imaging fidelity ${{\mathscr{F}}}_{{\rm{I}},1}=98.7(1)$% and ${{\mathscr{F}}}_{{\rm{I}},2}=98.2(2)$% per atom for charge-only and full-spin-charge resolution, respectively. The pumping fidelity is estimated by comparing the average number of atoms detected after pumping to the average number of atoms before pumping, and we obtain an average pumping fidelity of ${{\mathcal{F}}}_{{\rm{P}}}=97.6(1)$%, taking into account the slight discrepancy between ${{\mathscr{F}}}_{{\rm{I}},}$ and ${{\mathscr{F}}}_{{\rm{I}},{\rm{b}}}$. We deduce an overall detection fidelity of ${{\mathscr{F}}}_{1}={{\mathscr{F}}}_{{\rm{I}},1}=98.7(1)$% and ${{\mathscr{F}}}_{2}={{\mathscr{F}}}_{{\rm{I}},2}{{\mathscr{F}}}_{{\rm{P}}}=95.8(1)$% in the case of charge-only and full-spin-charge resolution, respectively.

Data statistics

We have taken approximately 19,000 experimental shots, iterating between mixD $\Delta \approx U/2$ and standard $\Delta = 0$. Here 61% of the shots have charge-only resolution and 39% have full spin and charge resolution.

The ladders are very sensitive to small drifts in the DMD pattern relative to the lattice sites. We thus keep track of the ladder potential by continuous automatic evaluation of the charge distribution and automatic feedback to the DMD pattern. If the average leg-to-leg occupation imbalance of standard ladders exceeds two holes, we dismiss the respective set of data due to the uncontrolled drift in the potential. For data analysis we then only take into

account ladders without double occupancies and with a leg-to-leg occupation imbalance of maximally one hole. This leaves us with more than 24,000 individual ladders, about half of which contain between two and four holes (Extended Data Fig. 4a). Most ladders show a magnetization $|M^z| < 2$, with ${M}^{z}={\sum }_{{\boldsymbol{i}}}{\hat{S}}_{{\boldsymbol{i}}}^{z}$ (Extended Data Fig. 4b). Figures and values given in the main text, unless otherwise mentioned, are filtered for two to four holes.

Numerical simulations using DMRG

We numerically simulate the t–J model, equation (1) in the main text, using MPS. For the mixD ($t_{\perp} = 0$) case, we set the parameters to $J_{\parallel}/J_{\perp} = 0.047$, $t_{\parallel}/J_{\perp} = 0.7$. In the standard ($t_{\perp} > 0$) case, the parameters are $J_{\parallel}/J_{\perp} = 0.06$, $t_{\parallel}/J_{\perp} = 0.9$ and $t_{\perp}/J_{\perp} = 3.57$. This corresponds to the t–J model derived from a Fermi–Hubbard model with $U/t_{\perp} = 14.16$, $t_{\parallel}/t_{\perp} = 0.26$ and, in the mixD case, $\Delta/U = 0.5$. We use the TeNPy package[51,52] to perform the MPS simulations. To simulate systems at finite temperature, we use the purification method[53,54], in which the Hilbert space is enlarged by an auxiliary site $a(i)$ per physical site i. The finite-temperature state of the physical system is obtained by tracing out the auxiliary degrees of freedom. We start from an infinite-temperature state, in which the physical and auxiliary degrees of freedom on each site are maximally entangled. In particular, we implement an entangler Hamiltonian[55] to prepare the infinite-temperature state of the t–J model. We work in the grand canonical ensemble and thus introduce a chemical potential μ to control the average number of holes in the system. Starting from the infinite-temperature state, we then use the W^{II}-time-evolution method[56] to perform imaginary time evolution up to the desired temperature. Depending on the system size, model (standard t–J versus mixD or Fermi–Hubbard), doping and temperature (finite temperature versus ground state), we use a bond dimension between $\chi = 50$ and $\chi = 400$. For the finite-temperature calculations, we use an imaginary time step of $\mathrm{d}t/J_{\perp} = 0.025$. We have carefully checked our results for convergence in the bond dimension and the size of the time step. We have benchmarked the MPS calculations by

comparing with exact diagonalization for small system sizes and find the same results.

To directly compare with the experimental data, we sample snapshots from the MPS using the perfect sampling algorithm[57]. In the evaluation of the snapshots, we account for the experimental detection fidelity by randomly placing artificial holes in the MPS snapshots according to our detection fidelity. We then apply the same filters regarding hole number and occupation imbalance as for the experimental data and model the hole number distribution of the experimental data (Extended Data Fig. 4a) by weighting the snapshots accordingly.

For ground-state simulations, for example to obtain the binding energies, we use the DMRG algorithm and work in a fixed ${S}_{z}^{{\rm{tot}}}$ and particle number sector.

From the Fermi–Hubbard to the *t*–*J* model

The Fermi–Hubbard model

$${\mathcal{H}}=-\,\sum _{\langle i\,,j\rangle ,\sigma }-{\widetilde{t}}_{ij}\,({\hat{c}}_{i\,,\sigma }^{\dagger }{\hat{c}}_{j,\sigma }+{\rm{h.c.}})\,+U\sum _{i}{\hat{n}}_{i,\uparrow }{\hat{n}}_{i,\downarrow }$$

contains a hopping term and (repulsive) on-site interaction. An additional potential offset $\varDelta$ on one of the two legs leads to

$${{\mathcal{H}}}_{\varDelta }={\mathcal{H}}\,+\varDelta \sum _{i\in (x,y=B)}{\hat{n}}_{i},$$

which cannot be generally reduced to an effective Hamiltonian with a tunnelling ${t}_{\perp }({\widetilde{t}}_{\perp },\varDelta)$, because in general the physics will depend both on the underlying Fermi–Hubbard tunnelling amplitude ${\widetilde{t}}_{\perp }$ and the offset $\varDelta$. An effective description only exists in the regime ${\widetilde{t}}_{\perp }\ll \varDelta \ll U$ (as well as for $U \ll \varDelta$ and the trivial $\varDelta = 0$), where ${\widetilde{t}}_{\perp }$ is eliminated from the Hamiltonian by working

in a time-dependent basis. We mention that, even in this regime, the effective model does not capture the full physics, but only holds for intermediate timescales for which the system is in a metastable state. For small tilts $| \varDelta | \ll | {\widetilde{t}}_{\perp }|$, no such metastability exists but instead the system directly equilibrates to a state in which more holes are in the upper leg. Such a system is not described by an effective Hamiltonian with mixed dimensionality, but by the full Hubbard model with $\varDelta$ and ${\widetilde{t}}_{\perp }$ terms.

In the limit of large interactions $U \gg t$, where $U/\widetilde{t}$ needs to be large enough to be well into the Mott-insulating regime, double occupancies are suppressed. An expansion to leading order in $\widetilde{t}/U$ yields several terms, including the t–J model of equation (1) with $J=4{\widetilde{t}}^{2}/U$ and thus $t \gg J$. In addition, the expansion yields terms of the order of t^2/U (ref. [58]) describing next-nearest-neighbour hopping by a (virtual) double occupancy, in analogy with the spin-exchange term.

For our mixD system the only term arising is approximately ${t}_{\parallel }^{2}/U\ll {J}_{\perp },{t}_{\parallel }$, which is much smaller than the relevant energy scales in the system and can thus be omitted. For the standard system there are more possible combinations of processes, such as approximately $t_{\parallel}t_{\perp}/U$, which is much larger than the process including only $t_{\parallel}$, but the system is still dominated by $t_{\perp}$, $t_{\parallel}$ and $J_{\perp}$. In the parameter regime in which $t_{\parallel} \gg J_{\perp}$, however, this term becomes increasingly important such that the Fermi–Hubbard system can eventually not be approximated by the t–J model of equation (1). This explains discrepancies found in the literature between binding energies calculated in t–J ladders[25,34] and in Fermi–Hubbard ladders[59] in the same parameter regimes.

Temperature estimation

We estimate the temperature of our system by comparing the measured rung spin correlations $C(0, 1)$ as defined in equation (2) to the values calculated from MPS snapshots (Extended Data Fig. 5a). We find that our average rung

spin correlations of $C(0, 1) = -0.38(1)$ for two to four holes correspond to a temperature of $k_{\rm B}T = 0.77(2)\,J_{\perp}$.

Our data are, however, not well described by a single spin correlation value, as we see variations both in time and across the four simultaneously realized ladders. The temperature estimation for the full dataset is therefore an average, and the data can contain features of both lower and higher temperatures. One reason for temperature variations are drifts in the apparatus on a timescale of days, affecting in particular the evaporation stage, which sets the global temperature. Another reason is the potential shaping, which distributes entropy between the four ladders and the surrounding bath. We thus attribute a temperature to each ladder (out of the four ladders we realize simultaneously) and each point in time by averaging the spin correlations of a time window of about ±12 h. The resulting spin correlation and temperature distribution are shown in Extended Data Fig. 5b.

Correlation functions

Evaluating correlators in finite-sized systems with fixed particle number leads to finite-size offsets, due to self-correlation of the particles. For our purpose we have to distinguish two cases. For correlations between different legs, for example, the rung hole correlation \({g}_{{\rm{h}}}^{(2)}(0,1)\), self-correlation does not cause problems. In the mixD case holes cannot move from one leg to the other, such that finding a hole in leg *A* does not influence the number of holes in leg *B*. In the standard case holes are mobile between the legs, but the focus of the analysis still lies on holes in opposite legs, because we select the data for low occupation imbalance. The correlations are thus not influenced by self-correlation. Correlations within the same leg are, however, strongly affected by finite-size offsets. We correct for these offsets using

$${g}_{{\rm{h}}}^{(2)}(d,0)=\frac{1}{{{\mathscr{N}}}_{d}}\sum _{{\boldsymbol{i}}-{\boldsymbol{j}}=(d,0)}\left(\frac{\langle {\hat{n}}_{{\boldsymbol{i}}}^{{\rm{h}}} {\hat{n}}_{{\boldsymbol{j}}}^{{\rm{h}}}\rangle }{\langle {\hat{n}}_{{\boldsymbol{i}}}^{{\rm{h}}}\rangle \langle

{\hat{n}}_{{\boldsymbol{j}}}^{{\rm{h}}}\rangle }\frac{{N}_{{\rm{l}}}}{{N}_{{\rm{l}}}-1}-\frac{L}{L-1}\right),$$

where $N_{\rm l}$ is the number of holes in the leg and L is the length of the leg. The same offset correction is applied to the pair correlator of equation (3) in the main text. The offset correction applies a distance independent correction and thus affects the overall value, but not the shape of the curve.

Binding energy

We estimate the binding energy of holes from the measured correlation \({g}_{{\rm{h}}}^{(2)}(0,1)\) of two holes on the same rung. To this end, we simplify the mixD t–J Hamiltonian of equation (1) by neglecting the two smallest energy scales $J_{\parallel}$, $t_{\parallel}$. This is partly justified by the fact that both are below the estimated temperature T of the experiment.

As a result, the Hamiltonian completely decouples into individual rungs and we can exactly diagonalize the latter. Then, as detailed below, we perform a canonical calculation of the entire system, with exactly one hole on each of the two legs of length L. From the known temperature T and the rung superexchange $J_{\perp}$ we obtain a direct relation between the binding energy $E_{\rm b}$ and the rung-correlation function \({g}_{{\rm{h}}}^{(2)}(0,1)\):

$${E}_{{\rm{b}}}=-{\beta }^{-1}\ln\left[\frac{\left(1+3{{\rm{e}}}^{-\beta {J}_{\perp }}\right)\left(1-\frac{{g}_{{\rm{h}}}^{(2)}(0,1)}{L-1}\right)}{4\left(1+{g}_{{\rm{h}}}^{(2)}(0,1)\right)}\right]$$
(4)

where $\beta = 1/(k_{\rm B}T)$.

To use the measured correlation value given in Fig. 2a, we have to eliminate the density dependence of the \({g}_{{\rm{h}}}^{(2)}\) correlator. Using the insights of Fig. 3b, we scale the hole correlator with the hole density $n_{\rm h}$. Using the scaled correlator \({g}_{{\rm{h}}}^{(2)}{n}_{{\rm{h}}}\) and the above formula with the experimentally estimated values for $k_{\rm B}T/J_{\perp}$ = 0.77(2), we obtain the estimate for the binding energy $E_{\rm b} = 0.82(6)\,J_{\perp}$ stated in the main text. The error derives from the error on the experimental value

and the error on the temperature estimation. If we use the measured hole correlation for exactly two holes in the system (Fig. 3b), as is used in the above derivation of (4), we obtain a binding energy of $E_b = 0.79(9)\,J_\perp$. Both calculations yield results in very good agreement with the theoretical prediction from DMRG at $L = 7$ of ${E}_{{\rm{b}}}^{{\rm{theo}}}=0.81\,{J}_{\perp }$. We calculate the binding energy in large systems using DMRG and find that the value settles quickly to around $E_{b,\infty} = 0.78\,J_\perp$. For length $L = 40$ rungs we find $E_{b,40} = 0.7805\,J_\perp$ and for $L = 80$ rungs we find $E_{b,80} = 0.7797\,J_\perp$. This demonstrates that our system with its tightly bound pairs provides a good approximation to the physics in larger systems.

In the remainder of this section, we explain the simplified model used here in more detail and derive from it equation (4). As mentioned in the beginning, we neglect the smallest energy scales $t_\parallel$ and $J_\parallel$. The eigenstates of each decoupled rung therefore become the two-hole state $| {\rm{hh}}\rangle$, the four spin-hole states $| {\rm{sh}},y,\sigma \rangle$ with leg index $y = 0, 1$ and spin index $\sigma = \uparrow,\downarrow$, the spin-singlet state $\left| {\rm{S}}\right\rangle$ and the three spin-triplet states $| {\rm{T}},m\rangle$ with $m = -1, 0, 1$. The corresponding eigenenergies are $\epsilon_{hh} = V$, $\epsilon_{sh} = \epsilon_T = 0$ and $\epsilon_S = -J_\perp$. Note that we allowed for a variable energy V of the hh state. For $t_\parallel = J_\parallel = 0$ we know that $V = 0$; however, for small but non-zero couplings $t_\parallel, J_\parallel$, a non-zero renormalization of $V \neq 0$ can be expected. The strength of V can be calculated perturbatively[22], but we treat it as a free parameter here, which allows us to go beyond a perturbative analysis.

We start by defining the binding energy of the simplified model in the thermodynamic limit $L \to \infty$. To this end, we compare the ground-state energy of a system with two independent holes, $2(E_{1h} - E_{0h}) = 2J_\perp$, with the ground-state energy of a system with one pair of bound holes, $E_{2h} - E_{0h} = V + J_\perp$; both are measured relative to the undoped ground state, $E_{0h} = -LJ_\perp$. The binding energy is then defined as

$${E}_{{\rm{b}}}=2{E}_{{\rm{1h}}}-{E}_{{\rm{0h}}}-{E}_{{\rm{2h}}}={J}_{\perp }-V.$$

For $E_{\rm b} > 0$ ($E_{\rm b} < 0$) the two-hole ground state is paired (unpaired).

To derive equation (4) we perform a canonical calculation with exactly one hole per leg. The probability of finding both holes on the same rung anywhere in the system becomes ${p}_{{\rm{hh}}}=L{{\rm{e}}}^{-\beta {E}_{{\rm{hh}}}}{Z}_{{\rm{S}}}^{L-1}/Z$, where we defined the spin ${Z}_{{\rm{S}}}={{\rm{e}}}^{-\beta {E}_{{\rm{S}}}}+3{{\rm{e}}}^{-\beta {E}_{{\rm{T}}}}$ and total partition functions $Z=L{{\rm{e}}}^{-\beta {E}_{{\rm{hh}}}}{Z}_{{\rm{S}}}^{L-1}+4L(L-1){{\rm{e}}}^{-2\beta {E}_{{\rm{sh}}}}{Z}_{{\rm{S}}}^{L-2}$. By the definition of the $g^{(2)}$ function provided in the main text, we obtain the relation

$${g}_{{\rm{h}}}^{(2)}(0,1)=\frac{{p}_{{\rm{hh}}}/L}{{(1/L)}^{2}}-1 \quad (5)$$

in our model by assuming a homogeneous density of $\langle {\hat{n}}_{i}\rangle =2/(2L)$ on each site and a homogeneous probability for an existing pair to occupy a specific rung of $1/L$. Thus $\langle {\hat{n}}_{i}^{{\rm{h}}}{\hat{n}}_{j}^{{\rm{h}}}\rangle = {p}_{{\rm{hh}}}/L$ for fixed (i, j) on one rung. There are thus L identical terms in the sum of ${g}_{{\rm{h}}}^{(2)}$ and one arrives at equation (5) by inserting ${{\mathcal{N}}}_{(0,1)}=L$. Simplifying this expression and solving for $E_{\rm b}$ finally yields equation (4).

Data availability

Source data are provided with this paper. The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Code availability

The code used for the analysis are available from the corresponding author on reasonable request.

References

1. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Microscopic theory of superconductivity. *Phys. Rev.* **106**, 162–164 (1957).

2. Scalapino, D. J. Superconductivity and spin fluctuations. *J. Low Temp. Phys.* **117**, 179–188 (1999).

3. Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78**, 17–85 (2006).

4. Dagotto, E., Riera, J. & Scalapino, D. Superconductivity in ladders and coupled planes. *Phys. Rev. B* **45**, 5744–5747 (1992).

5. Sigrist, M., Rice, T. M. & Zhang, F. C. Superconductivity in a quasi-one-dimensional spin liquid. *Phys. Rev. B* **49**, 12058–12061 (1994).

6. Troyer, M., Tsunetsugu, H. & Rice, T. M. Properties of lightly doped t-J two-leg ladders. *Phys. Rev. B* **53**, 251–267 (1996).

7. Schulz, H. J. Metal-insulator transition in the two-chain model of correlated fermions. *Phys. Rev. B* **59**, R2471–R2473 (1999).

8. Giamarchi, T. *Quantum Physics in One Dimension*. International Series of Monographs on Physics, vol. 121 (Clarendon Press, 2004).

9. Kohn, W. & Luttinger, J. M. New mechanism for superconductivity. *Phys. Rev. Lett.* **15**, 524–526 (1965).

10. Bohrdt, A., Homeier, L., Bloch, I., Demler, E. & Grusdt, F. Strong pairing in mixed dimensional bilayer antiferromagnetic Mott insulators. *Nat. Phys.* **18**, 651–656 (2021).

11. White, B., Thompson, J. & Maple, M. Unconventional superconductivity in heavy-fermion compounds. *Phys. C: Supercond. Appl.* **514**, 246–278 (2015).

12. Wen, H.-H. & Li, S. Materials and novel superconductivity in iron pnictide superconductors. *Annu. Rev. Condens. Matter Phys.* **2**, 121–140 (2011).

13. Wosnitza, J. Superconductivity in layered organic metals. *Crystals* **2**, 248–265 (2012).

14. Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S. & Zaanen, J. From quantum matter to high-temperature superconductivity in copper oxides. *Nature* **518**, 179–186 (2015).

15. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).

16. Gross, C. & Bloch, I. Quantum simulations with ultracold atoms in optical lattices. *Science* **357**, 995–1001 (2017).

17. Chiu, C. S. et al. String patterns in the doped Hubbard model. *Science* **365**, 251–256 (2019).

18. Brown, P. T. et al. Bad metallic transport in a cold atom Fermi-Hubbard system. *Science* **363**, 379–382 (2019).

19. Koepsell, J. et al. Imaging magnetic polarons in the doped Fermi-Hubbard model. *Nature* **572**, 358–362 (2019).

20. Ji, G. et al. Coupling a mobile hole to an antiferromagnetic spin background: transient dynamics of a magnetic polaron. *Phys. Rev. X* **11**, 021022 (2021).

21. Koepsell, J. et al. Microscopic evolution of doped Mott insulators from polaronic metal to Fermi liquid. *Science* **374**, 82–86 (2021).

22. Bohrdt, A., Homeier, L., Reinmoser, C., Demler, E. & Grusdt, F. Exploration of doped quantum magnets with ultracold atoms. *Ann. Phys.* **435**, 168651 (2021).

23. Qin, M. et al. Absence of superconductivity in the pure two-dimensional Hubbard model. *Phys. Rev. X* **10**, 031016 (2020).

24. Arrigoni, E., Fradkin, E. & Kivelson, S. A. Mechanism of high-temperature superconductivity in a striped Hubbard model. *Phys. Rev. B* **69**, 214519 (2004).

25. Chernyshev, A. L., Leung, P. W. & Gooding, R. J. Comprehensive numerical and analytical study of two holes doped into the two-dimensional $t-J$ model. *Phys. Rev. B* **58**, 13594–13613 (1998).

26. White, S. R. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69**, 2863–2866 (1992).

27. Uehara, M. et al. Superconductivity in the ladder material $Sr_{0.4}Ca_{13.6}Cu_{24}O_{41.84}$. *J. Phys. Soc. Jpn.* **65**, 2764–2767 (1996).

28. Nagata, T. et al. Pressure-induced dimensional crossover and superconductivity in the hole-doped two-leg ladder compound $Sr_{14-x}Ca_xCu_{24}O_{41}$. *Phys. Rev. Lett.* **81**, 1090–1093 (1998).

29. Dagotto, E. Experiments on ladders reveal a complex interplay between a spin-gapped normal state and superconductivity. *Rep. Prog. Phys.* **62**, 1525–1571 (1999).

30. Chakravarty, S. & Kivelson, S. A. Electronic mechanism of superconductivity in the cuprates, C_{60}, and polyacenes. *Phys. Rev. B* **64**, 064511 (2001).

31. Kantian, A., Dolfi, M., Troyer, M. & Giamarchi, T. Understanding repulsively mediated superconductivity of correlated electrons via massively parallel density matrix renormalization group. *Phys. Rev. B* **100**, 075138 (2019).

32. Leung, P. W. Low-energy states with different symmetries in the $t-J$ model with two holes on a 32-site lattice. *Phys. Rev. B* **65**, 205101 (2002).

33. Blomquist, E. & Carlström, J. Evidence of attraction between charge carriers in a doped Mott insulator. *Phys. Rev. Res.* **3**, 013272 (2021).

34. White, S. R. & Scalapino, D. J. Hole and pair structures in the t-J model. *Phys. Rev. B* **55**, 6504–6517 (1997).

35. Duan, L.-M., Demler, E. & Lukin, M. D. Controlling spin exchange interactions of ultracold atoms in optical lattices. *Phys. Rev. Lett.* **91**, 090402 (2003).

36. Trotzky, S. et al. Time-resolved observation and control of superexchange interactions with ultracold atoms in optical lattices. *Science* **319**, 295–299 (2008).

37. Sompet, P. et al. Realizing the symmetry-protected haldane phase in Fermi-Hubbard ladders. *Nature* **606**, 484–488 (2020).

38. Koepsell, J. et al. Robust bilayer charge pumping for spin- and density-resolved quantum gas microscopy. *Phys. Rev. Lett.* **125**, 010403 (2020).

39. Fradkin, E., Kivelson, S. A. & Tranquada, J. M. Colloquium: theory of intertwined orders in high temperature superconductors. *Rev. Mod. Phys.* **87**, 457–482 (2015).

40. Friedel, J. The distribution of electrons round impurities in monovalent metals. *Philos. Mag. J. Sci.* **43**, 153–189 (1952).

41. White, S. R., Affleck, I. & Scalapino, D. J. Friedel oscillations and charge density waves in chains and ladders. *Phys. Rev. B* **65**, 165122 (2002).

42. Zhang, S.-S., Zhu, W. & Batista, C. D. Pairing from strong repulsion in triangular lattice Hubbard model. *Phys. Rev. B* **97**, 140507 (2018).

43. Song, Z.-D. & Bernevig, B. A. MATBG as topological heavy fermion: I. exact mapping and correlated insulators. Preprint at https://arxiv.org/abs/2111.05865 (2021).

44. Chaudhary, S., Hsieh, D. & Refael, G. Orbital floquet engineering of exchange interactions in magnetic materials. *Phys. Rev. B* **100**, 220403 (2019).

45. Preiss, P. M., Ma, R., Tai, M. E., Simon, J. & Greiner, M. Quantum gas microscopy with spin, atom-number, and multilayer readout. *Phys. Rev. A* **91**, 041602 (2015).

46. Hartke, T., Oreg, B., Jia, N. & Zwierlein, M. Doublon-hole correlations and fluctuation thermometry in a Fermi-Hubbard gas. *Phys. Rev. Lett.* **125**, 113601 (2020).

47. Grusdt, F., Zhu, Z., Shi, T. & Demler, E. Meson formation in mixed-dimensional t-J models. *SciPost Phys.* **5**, 057 (2018).

48. Rey, A. M. et al. Controlled preparation and detection of d-wave superfluidity in two-dimensional optical superlattices. *Europhys. Lett.* **87**, 60001 (2009).

49. Trebst, S., Schollwöck, U., Troyer, M. & Zoller, P. *d*-Wave resonating valence bond states of fermionic atoms in optical lattices. *Phys. Rev. Lett.* **96**, 250402 (2006).

50. Omran, A. et al. Microscopic observation of Pauli blocking in degenerate Fermionic lattice gases. *Phys. Rev. Lett.* **115**, 263001 (2015).

51. Hauschild, J. et al. Tensor network python. *GitHub* https://github.com/tenpy/tenpy/ (2018).

52. Hauschild, J. & Pollmann, F. Efficient numerical simulations with tensor networks: Tensor Network Python (TeNPy). *SciPost Phys. Lect. Notes 5* (2018).

53. Verstraete, F., García-Ripoll, J. J. & Cirac, J. I. Matrix product density operators: simulation of finite-temperature and dissipative systems. *Phys. Rev. Lett.* **93**, 207204 (2004).

54. Schollwöck, U. The density-matrix renormalization group in the age of matrix product states. *Ann. Phys.* **326**, 96–192 (2011).

55. Nocera, A. & Alvarez, G. Symmetry-conserving purification of quantum states within the density matrix renormalization group. *Phys.*

Rev. B **93**, 045137 (2016).

56. Zaletel, M. P., Mong, R. S. K., Karrasch, C., Moore, J. E. & Pollmann, F. Time-evolving a matrix product state with long-ranged interactions. *Phys. Rev. B* **91**, 165112 (2015).

57. Ferris, A. J. & Vidal, G. Perfect sampling with unitary tensor networks. *Phys. Rev. B* **85**, 165146 (2012).

58. Auerbach, A. *Interacting Electrons and Quantum Magnetism* Graduate Texts in Contemporary Physics (Springer, 1994).

59. Karakonstantakis, G., Berg, E., White, S. R. & Kivelson, S. A. Enhanced pairing in the checkerboard Hubbard ladder. *Phys. Rev. B* **83**, 054508 (2011).

Acknowledgements

We thank T. Giamarchi for discussions. We thank A. Kantian for discussions and critical reading of the manuscript. This work was supported by the Max Planck Society (MPG), the European Union (FET-Flag 817482, PASQUANS), the Max Planck Harvard Research Center for Quantum Optics (MPHQ), the German Federal Ministry of Education and Research (BMBF grant agreement 13N15890, FermiQP) and Germany's Excellence Strategy (EXC-2111-390814868). T.C. acknowledges funding from the Alexander v. Humboldt Foundation. F.G. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 948141) from ERC Starting Grant SimUcQuam. A.B. acknowledges funding from the National Science Foundation (NSF) through a grant for the Institute for Theoretical Atomic, Molecular, and Optical Physics at Harvard University and the Smithsonian Astrophysical Observatory. E.D. acknowledges funding from the Army Research Office (ARO) (grant number W911NF-20-1-0163) and AFOSR-MURI: Photonic Quantum Matter award FA95501610323.

Funding

Open access funding provided by Max Planck Society.

Author information

Authors and Affiliations

1. Max-Planck-Institut für Quantenoptik, Garching, Germany

   Sarah Hirthe, Thomas Chalopin, Dominik Bourgund, Petar Bojović, Immanuel Bloch & Timon A. Hilker

2. Munich Center for Quantum Science and Technology, Munich, Germany

   Sarah Hirthe, Thomas Chalopin, Dominik Bourgund, Petar Bojović, Fabian Grusdt, Immanuel Bloch & Timon A. Hilker

3. Department of Physics, Harvard University, Cambridge, MA, USA

   Annabelle Bohrdt

4. ITAMP, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA

   Annabelle Bohrdt

5. Institute for Theoretical Physics, ETH Zurich, Zurich, Switzerland

   Eugene Demler

6. Department of Physics and Arnold Sommerfeld Center for Theoretical Physics (ASC), Ludwig-Maximilians-Universität, Munich, Germany

   Fabian Grusdt

7. Department of Physics, Ludwig-Maximilians-Universität, Munich, Germany

Fabian Grusdt & Immanuel Bloch

Contributions

S.H. led the project. S.H., T.C., P.B. and D.B. contributed significantly to data collection and analysis. A.B. carried out the numeric calculations. S.H., T.C. and T.A.H. wrote the manuscript. E.D., F.G., I.B. and T.A.H supervised the study. All authors contributed extensively to interpretation of the data and production of the manuscript.

Corresponding author

Correspondence to Sarah Hirthe.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Randall Hulet for their contribution to the peer review of this paper. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Preparation sequence for mixD systems.

a, We first prepare nearly uncoupled 1D chains in which the leg tunnelling exceeds the rung coupling. **b**, While the legs are decoupled, we apply the offset Δ to one leg of the ladder. **c**, The final parameters are reached by ramping down the leg coupling and ramping up the rung coupling. There, the potential offset Δ between legs prevents tunnelling from one leg to the other. Note that in the final configuration $J_\perp \gg J_\parallel$.

Extended Data Fig. 2 Calibration of the optical potential offset.

The experimental sequence is run for different value of Δ in a regime close to unit occupancy of the lattice. tunnelling from one leg to the other is suppressed as long as $|\Delta - U| > 0$. When $\Delta \sim U$, tunnelling is possible, and an increased number of doublons in the system is measured.

Extended Data Fig. 3 Density of the mixD system without doublons.

The density of the mixD system, where only ladders without double occupancies are taken into account.

Extended Data Fig. 4 Hole and magnetization statistics.

a,**b**, Experimental distribution of holes per ladder (**a**) and total magnetization (**b**) for the data shown in Fig. 2a–c, and Fig. 3a.

Extended Data Fig. 5 Temperature estimation.

a, Singlet strength versus temperature. The calibration of temperature is performed using MPS data containing two to four holes. **b**, Experimental singlet strength and **c**, inferred temperature distributions. We evaluate our rung spin correlations $C(0, 1)$ on the mixD system, using a time window of about 24 h. The temperature is extracted from $C(0, 1)$ using the MPS simulation (**a**).

Supplementary information

Supplementary information

Supplementary Figs. 1–7 and references.

Peer Review File

Source data

Source Data Fig. 1

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 4

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Hirthe, S., Chalopin, T., Bourgund, D. *et al.* Magnetically mediated hole pairing in fermionic ladders of ultracold atoms. *Nature* **613**, 463–467 (2023). https://doi.org/10.1038/s41586-022-05437-y

- Received: 18 March 2022
- Accepted: 11 October 2022
- Published: 18 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05437-y

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

[Exotic pairing of charge carriers seen in a quantum simulation]

Nature Research Briefing 18 Jan 2023

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05437-y

- Article
- Open Access
- Published: 18 January 2023

Octupole-driven magnetoresistance in an antiferromagnetic tunnel junction

- Xianzhe Chen,
- Tomoya Higo,
- Katsuhiro Tanaka,
- Takuya Nomoto,
- Hanshen Tsai,
- Hiroshi Idzuchi,
- Masanobu Shiga,
- Shoya Sakamoto,
- Ryoya Ando,
- Hidetoshi Kosaki,
- Takumi Matsuo,
- Daisuke Nishio-Hamane,
- Ryotaro Arita,
- Shinji Miwa &
- …
- Satoru Nakatsuji

Show authors

Nature **volume 613**, pages 490–495 (2023)

- 2165 Accesses
- 87 Altmetric

- Metrics details

Abstract

The tunnelling electric current passing through a magnetic tunnel junction (MTJ) is strongly dependent on the relative orientation of magnetizations in ferromagnetic electrodes sandwiching an insulating barrier, rendering efficient readout of spintronics devices[1,2,3,4,5]. Thus, tunnelling magnetoresistance (TMR) is considered to be proportional to spin polarization at the interface[1] and, to date, has been studied primarily in ferromagnets. Here we report observation of TMR in an all-antiferromagnetic tunnel junction consisting of Mn_3Sn/MgO/Mn_3Sn (ref. [6]). We measured a TMR ratio of around 2% at room temperature, which arises between the parallel and antiparallel configurations of the cluster magnetic octupoles in the chiral antiferromagnetic state. Moreover, we carried out measurements using a Fe/MgO/Mn_3Sn MTJ and show that the sign and direction of anisotropic longitudinal spin-polarized current in the antiferromagnet[7] can be controlled by octupole direction. Strikingly, the TMR ratio (about 2%) of the all-antiferromagnetic MTJ is much larger than that estimated using the observed spin polarization. Theoretically, we found that the chiral antiferromagnetic MTJ may produce a substantially large TMR ratio as a result of the time-reversal, symmetry-breaking polarization characteristic of cluster magnetic octupoles. Our work lays the foundation for the development of ultrafast and efficient spintronic devices using antiferromagnets[8,9,10].

Main

The discovery of giant magnetoresistance[11] and tunnelling magnetoresistance (TMR)[1,2,3,4,5] has facilitated rapid progress in the highly integrated and efficient spintronic technologies based on ferromagnets. TMR and spin-transfer torque (STT), due to longitudinal spin-polarized current, provide the reading and writing protocols for the two-terminal magnetoresistive random access memory (MRAM) recently commercialized[12,13,14]. In addition to the well-established ferromagnetic

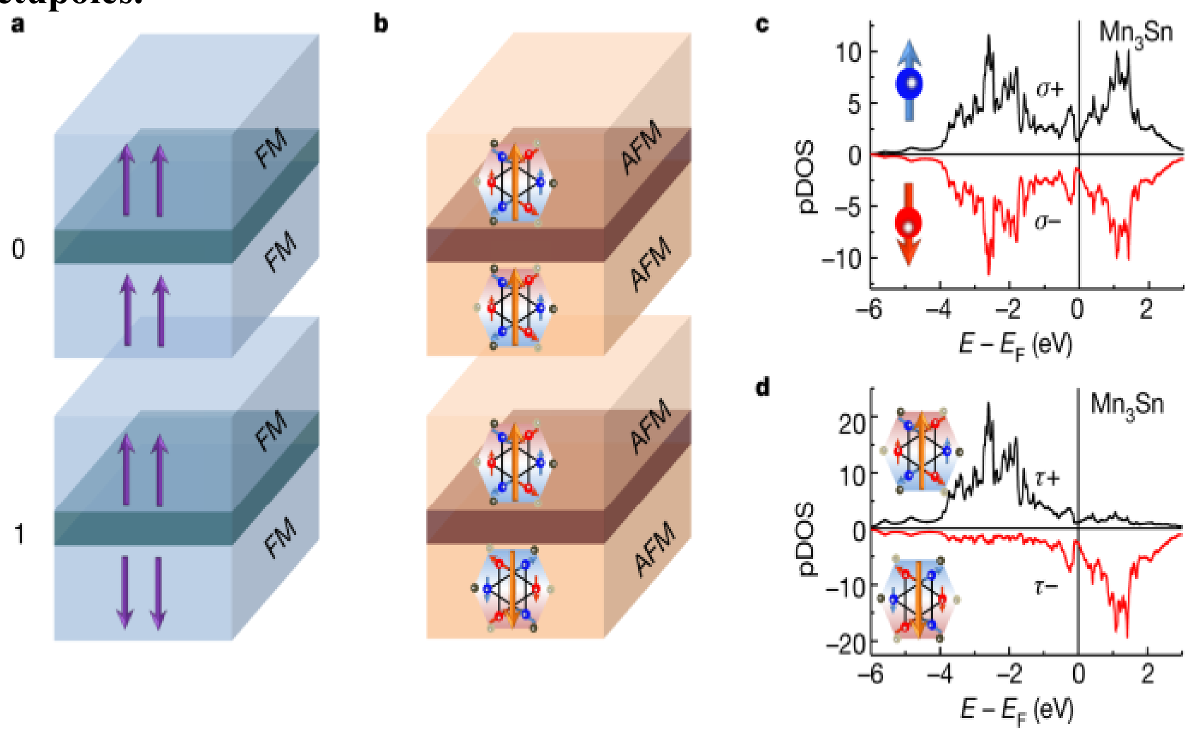
spintronics, antiferromagnets have attracted tremendous interest[6,7,8,9,10,15] as next-generation active elements for further improvement in operation speed and integration density. Whereas no STT method has been demonstrated experimentally, various writing means have been developed[16,17,18,19], including spin-orbit torque[20,21,22], using the same methods used for ferromagnets. In regard to readout, the anomalous Hall effect[6,23,24,25,26] has recently become available for antiferromagnets in addition to anisotropic magnetoresistance[8,9,10,18,27] and spin Hall magnetoresistance[19]. In addition, tunnelling anisotropic magnetoresistance (TAMR) has been investigated for junctions based on a single antiferromagnetic electrode[28,29,30]. However, TAMR is generally seen at low temperatures and those much lower than the TMR, even when it becomes available at room temperature.

Regarding the perspective of future applications, it is highly important to develop TMR using all-antiferromagnetic MTJs, which potentially generates a high magnetoresistance ratio. On the other hand, research and development of the TMR effect has been restricted primarily to ferromagnetic MTJs because the effect is considered to derive from finite spin polarization at the interface between the magnetic electrodes and insulating barrier[1,2,3,4,5]. Thus, no reports have been published to date on the TMR effect using an all-antiferromagnetic MTJ. In addition, as the basis for designing the STT–MRAM based solely on antiferromagnets, observation and manipulation of longitudinal spin-polarized current is necessary but, again, has never been performed for antiferromagnets. Here we report our discovery of a finite TMR of approximately 2% using a MTJ solely comprising an antiferromagnet. Moreover, our experiment has further clarified the existence of the anisotropic, longitudinal spin-polarized current in an antiferromagnet. Strikingly, the size of TMR in the all-antiferromagnetic MTJ is much larger than the conventional estimate based on spin polarization and thus should derive from a new mechanism.

Ferromagnetic TMR produces the binary signals '0' and '1', respectively, via parallel and antiparallel arrangements of spin polarization in a pair of ferromagnetic electrodes (Fig. 1a). In principle, analogous binary states can be defined for an antiferromagnetic MTJ providing that the antiferromagnetic states break time-reversal symmetry (TRS) macroscopically. Such antiferromagnets have recently been discovered and

intensively studied for the development of antiferromagnetic spintronics, because they produce large transverse effects despite vanishingly small magnetization[6,23,24,25,26,31,32]. Breaking of TRS in these antiferromagnetic states leads to characteristic electronic structures such as magnetic Weyl semimetal state[33,34], spin-splitting bands[7,31,32,35,36] and polarization of cluster magnetic multipoles[37].

Fig. 1: Antiferromagnetic tunnel junction with cluster magnetic octupoles.



a, Schematics of ferromagnetic (FM) tunnel junction sandwiching an insulating barrier (middle layer, darker colour); the parallel 0 and antiparallel 1 configurations of ferromagnetic moments (purple arrows) generate two different resistance states. **b**, Schematics of a tunnel junction made solely of antiferromagnetic (AFM) electrodes with a magnetic order parameter-breaking TRS, such as the cluster magnetic octupole in Mn_3Sn. The antiferromagnet Mn_3Sn has a hexagonal Ni_3Sn-type structure (space group P63/mmc)[6], and Mn moments (red and blue arrows) form a cluster magnetic octupole (orange arrows) consisting of six spins on the kagome bilayer. States 0 and 1 arise when the top and bottom octupole polarizations are parallel and antiparallel, respectively. **c**, Projected density of states

(pDOS) onto 3*d* orbitals computed using DFT for two opposite spin states with σ^{+} and σ^{-} in Mn_3Sn. The pDOS is symmetric in terms of spin polarization (red and blue arrows). Here the direction of spin polarization is parallel to the *a* axis of the crystal. **d**, Calculated pDOS with projection onto octupolar ordered states with opposite octupole moments τ+ and τ−. Here the polarization direction of the cluster magnetic octupole is parallel to the *a* axis of the crystal and lies in the kagome plane. Inset, corresponding octupole-majority and -minority states used for calculations. Black and red lines correspond to octupole-majority ($pDOS_{maj}$) and octupole-minority ($pDOS_{min}$) DOSs, respectively. The absolute value of $pDOS_{maj}$ is higher when energy level E is lower than Fermi energy E_F, whereas that of $pDOS_{min}$ dominates at higher E. The polarity of octupole polarization changes sign by tuning the Fermi energy of Mn_3Sn (Extended Data Fig. 2c).

Source data

Here we focus on a specific example, the chiral antiferromagnet Mn_3Sn. This hexagonal kagome metal has attracted much interest because it exhibits large transverse responses normally absent in antiferromagnets[6,38,39,40]. Below the Néel temperature $T_N \approx 430$ K, all Mn moments of approximately 3 μ_B lying in the kagome plane form an antichiral 120° spin order. The magnetic texture can be viewed as a ferroic order of cluster magnetic octupole lying in the kagome plane (Fig. 1b and Methods)[37,39]. This ferroic octupole order breaks the TRS macroscopically and allows us to define states 0 and 1 as the parallel and antiparallel configurations, respectively, of octupole polarization (Fig. 1b, orange arrows and Methods). In addition, it is ferroic octupole order that stabilizes the magnetic Weyl semimetal state[33,34] and drives large transverse responses such as the anomalous Hall effect[6], anomalous Nernst effect[38] and magneto-optical Kerr effect (MOKE)[39].

To further demonstrate the analogy between spin polarization in ferromagnets, and octupole polarization in antiferromagnets, we drew the projected density of states (DOS) for the positive and negative polarization of spins (Fig. 1c) and magnetic octupoles (Fig. 1d) in the chiral antiferromagnetic state of Mn_3Sn. In sharp contrast with the symmetric DOS projected onto spin-up and -down states (Fig. 1c), we found a clear energy

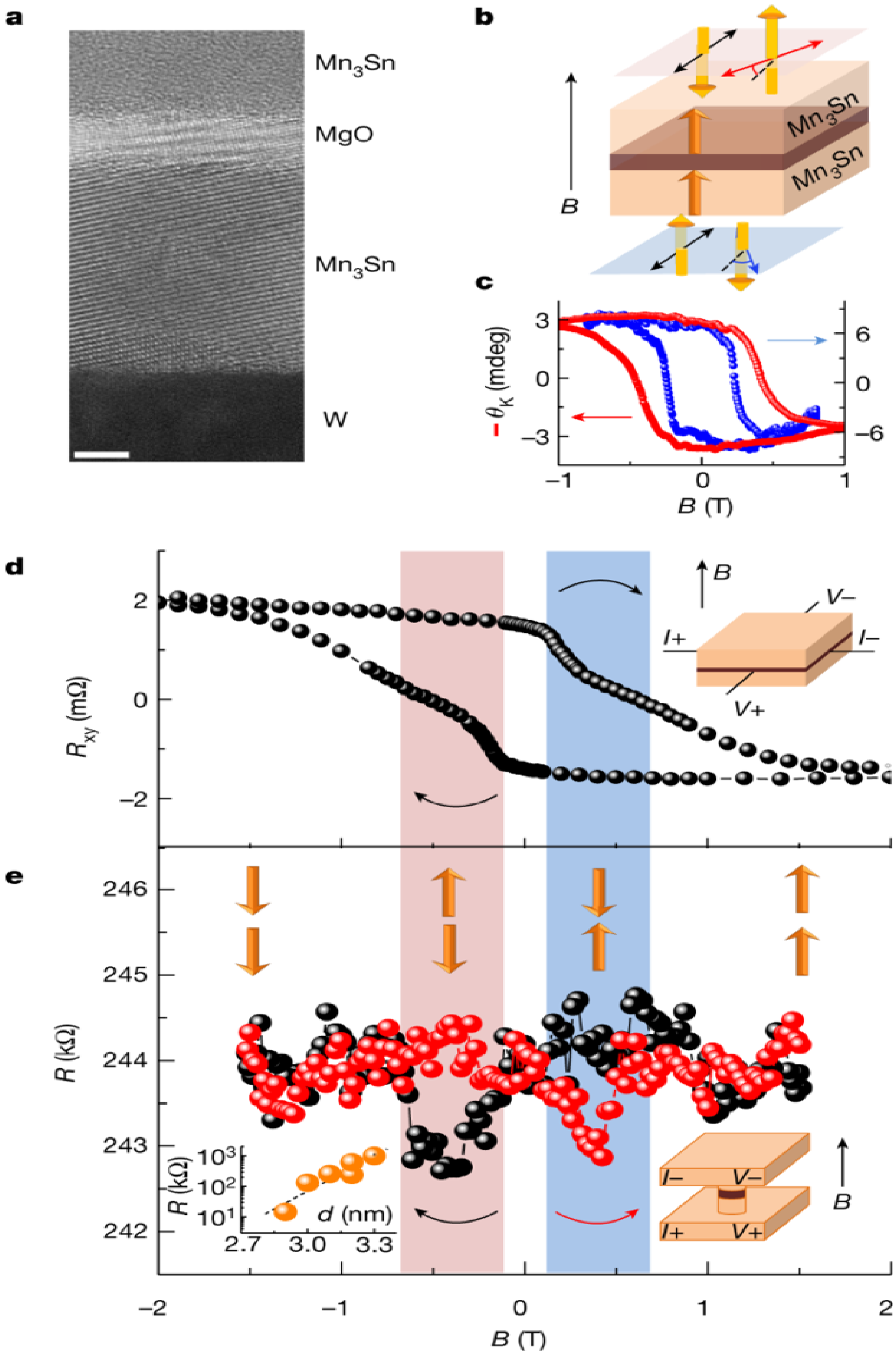
shift between the majority and minority bands of the octupole moment (Fig. 1d and Methods), closely resembling the shift in the majority and minority bands of the spin moment in ferromagnetic iron (Extended Data Fig. 2a). For conventional ferromagnetic TMR, the imbalance in spin polarization is the key origin according to the basic model for TMR proposed by Julliere[1] (Fig. 1a, Extended Data Fig. 2a,b and Methods). This suggests that similar TRS breaking due to the imbalance seen in octupole polarization might lead to antiferromagnetic TMR (Fig. 1b,d).

To examine whether a finite TMR effect can be generated between the parallel and antiparallel arrangements of the cluster magnetic octupole moments in Mn_3Sn, we carried out a theoretical simulation from first principles (Methods and Extended Data Fig. 3). We used specific geometry for calculations—namely, the Mn_3Sn/vacuum/Mn_3Sn (0001) structure, in which octupole moments are inside the plane perpendicular to the conducting path. We calculated tunnelling conductances for the parallel and antiparallel arrangements of the octupole moments at the Fermi level, which give the finite positive TMR ratio (Extended Data Fig. 3d and Methods). Although the net magnetic moments of Mn_3Sn were vanishingly small, the calculated TMR ratio was as large as that obtained in a ferromagnetic counterpart, Fe/vacuum/Fe MTJ (Extended Data Fig. 3i). This indicates that Mn_3Sn has great potential for use in TMR-based devices such as MRAM, because of the macroscopic TRS breaking captured by cluster octupole moments (Methods). In addition, we found that the transmission properties are robust against both interfacial disorder and the relative lateral shift of the atomic layer at the interface, being useful for TMR applications (Methods, Extended Data Fig. 5 and Supplementary Figs. 4 and 5).

Motivated by this possibility, we fabricated the Mn_3Sn-based MTJ to demonstrate the TMR effect corresponding to binary states 0 and 1. To observe reliable tunnelling conduction, preparation of a continuous and smooth tunnelling barrier is essential. As shown in Fig. 2a, we fabricated the stack consisting of W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm)/Mn_3Sn (42 nm)/Ta (5 nm) from the MgO substrate side (Methods). According to theory, tunnelling conductance can be potentially anisotropic (Supplementary Information and Supplementary Fig. 1). On the other hand, characterization of films by Hall and Kerr effects may be possible only when octupole

polarization has the out-of-plane component. Given these two constraints, we fabricated an epitaxial film of Mn_3Sn with the $(01\bar{1}1)$ orientation aligned close to the out-of-plane direction by taking advantage of the fact that the MgO barrier tends to grow along the [001] direction (Extended Data Fig. 10 and Methods). Transmission electron microscopy (TEM) imaging confirmed the presence of stack films with a sharp and continuous interface between Mn_3Sn and MgO (Fig. 2a, Extended Data Fig. 9 and Methods). In addition, the *c* axis of Mn_3Sn was found to be approximately 30° off the normal direction of the film (Extended Data Fig. 10 and Methods), allowing us to characterize the films and to measure tunnelling conduction in the Mn_3Sn/MgO/Mn_3Sn trilayer, as discussed below.

Fig. 2: Observation of room-temperature magnetoresistance in the antiferromagnetic tunnel junction based on the chiral antiferromagnet Mn_3Sn.



a
Mn3Sn
MgO
Mn3Sn
W
b
B
Mn3Sn
Mn3Sn
c
θK (mdeg)
θK (mdeg)
B (T)
d
Rxy (mΩ)
B
I+
I−
V−
V+
e
R (kΩ)
R (kΩ)
d (nm)
I−
V−
I+
V+
B
B (T)

a, High-resolution transmission electron microscopy image of the heterostructure consisting of the antiferromagnetic tunnel junction, W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm)/Mn_3Sn (42 nm)/Ta (5 nm), on the MgO substrate. Scale bar, 3 nm. Continuous layers and sharp interfaces were observed (Extended Data Figs. 9 and 10). **b**, Schematic illustration of MOKE measurements. Net polarization directions of magnetic octupoles are denoted by orange arrows. A polarized light beam was applied perpendicular to the film plane (from both the bottom and top), and reflected light became elliptically polarized with the major axis rotated by polar MOKE angle θ_K. **c**, Field dependence of θ_K for polarized light applied to the top plane (red) and bottom plane (blue) of Mn_3Sn. **d**, Hall resistance versus magnetic field, B applied perpendicular to Mn_3Sn planes and electric current, I. **e**. Field dependence of tunnelling resistance measured in magnetic field, B and electric current, I, both applied perpendicular to the Mn_3Sn plane. Four different octupole configurations were obtained, namely from the positive high-field side and up–up, down–up (blue region), up–down (red region) and down–down configurations. Here we show only octupole configurations, for visual clarity. Inset, tunnelling resistance R as a function of MgO thickness d. **d**,**e**, Schematics for measurement configurations are also shown in insets (Methods). The TMR sign is negative (Methods). In addition, our measurements of the inelastic electron tunnelling spectrum suggest that magnon-electron scattering in the energy range up to 0.1 eV may play a key role in TMR (Supplementary Information and Supplementary Fig. 3).

Source data

For observation of the TMR effect, the top and bottom Mn_3Sn electrodes are required to have two distinctive coercivities so that the parallel and antiparallel configurations of octupole polarizations arise as a function of magnetic field (Fig. 1b). We characterized the coercivities of the two Mn_3Sn layers in the MTJ by field-sweep measurements of polar MOKE at room temperature, where a polarized light beam was injected into both top and bottom layers (Fig. 2b and Methods). Figure 2c presents polar MOKE loops measured in magnetic field B applied perpendicular to the plane. The top and bottom Mn_3Sn layers produced clear hysteresis, with square loops as a function of field B, and yielded sizeable changes in Kerr rotation angles

$|\theta_K|$ (Fig. 2c). It should be noted that the coercive field of B_c, at approximately 0.5 T for the top Mn_3Sn (Fig. 2c, red), is twofold that of B_c at 0.25 T for the bottom layer (Fig. 2c, blue). Therefore, these two coercivities should generate state 1 when top and bottom octupole moments are antiparallel.

We further examined the formation of binary states via Hall measurements in a magnetic field perpendicular to the film plane. To stabilize state 0, in which octupole moments in both the top and bottom Mn_3Sn point upward, we first applied the magnetic field of +2 T, which is larger than all coercive fields. By sweeping the field from +2 to −2 T (Fig. 2d) we found two transitions at B_c (approximately −0.2 and approximately −0.6 T) close to the coercivities seen in the MOKE measurements above, supporting the premise that coercive fields are nearly identical in the bulk and interface detected by the Hall effect and MOKE measurements, respectively. When the field passed through the smaller coercivity at B_c (approximately –0.2 T), the octupole moment in the bottom Mn_3Sn layer reversed, producing state 1 with antiparallel configuration (red regions in Fig. 2d,e). When the field passed through the larger coercivity of B_c(approximately −0.6 T), the octupole moment in the top layer also began rotating downwards, establishing state 0 in which octupole moments were again parallel. When the field was scanned back from −2 to +2 T, these two transitions also occurred at B_c(approximately 0.2 and approximately 0.6 T), stabilizing a further state 1 between 0.2 and 0.6 T (blue regions in Fig. 2d,e), as well as a state 0 greater than 0.6 T.

We then moved our focus to the field dependence of the TMR in the antiferromagnetic MTJ. By scanning the field applied perpendicularly to the film plane from +1.5 to −1.5 T, we found that resistance exhibited a clear drop when B was between −0.2 and −0.6 T. This low-resistance state (LRS) appeared again between 0.2 and 0.6 T when we swept the field backwards from −1.5 to +1.5 T. The LRS and the remaining high-resistance state (HRS) correspond to states 1 and 0 in the MOKE and Hall measurements (Fig. 2c,d)—that is, the antiparallel and parallel configurations of the top and bottom octupole moments, respectively. By adapting the definition TMR = (HRS – LRS)/LRS × 100%, we estimated the TMR ratio to be around 0.6%

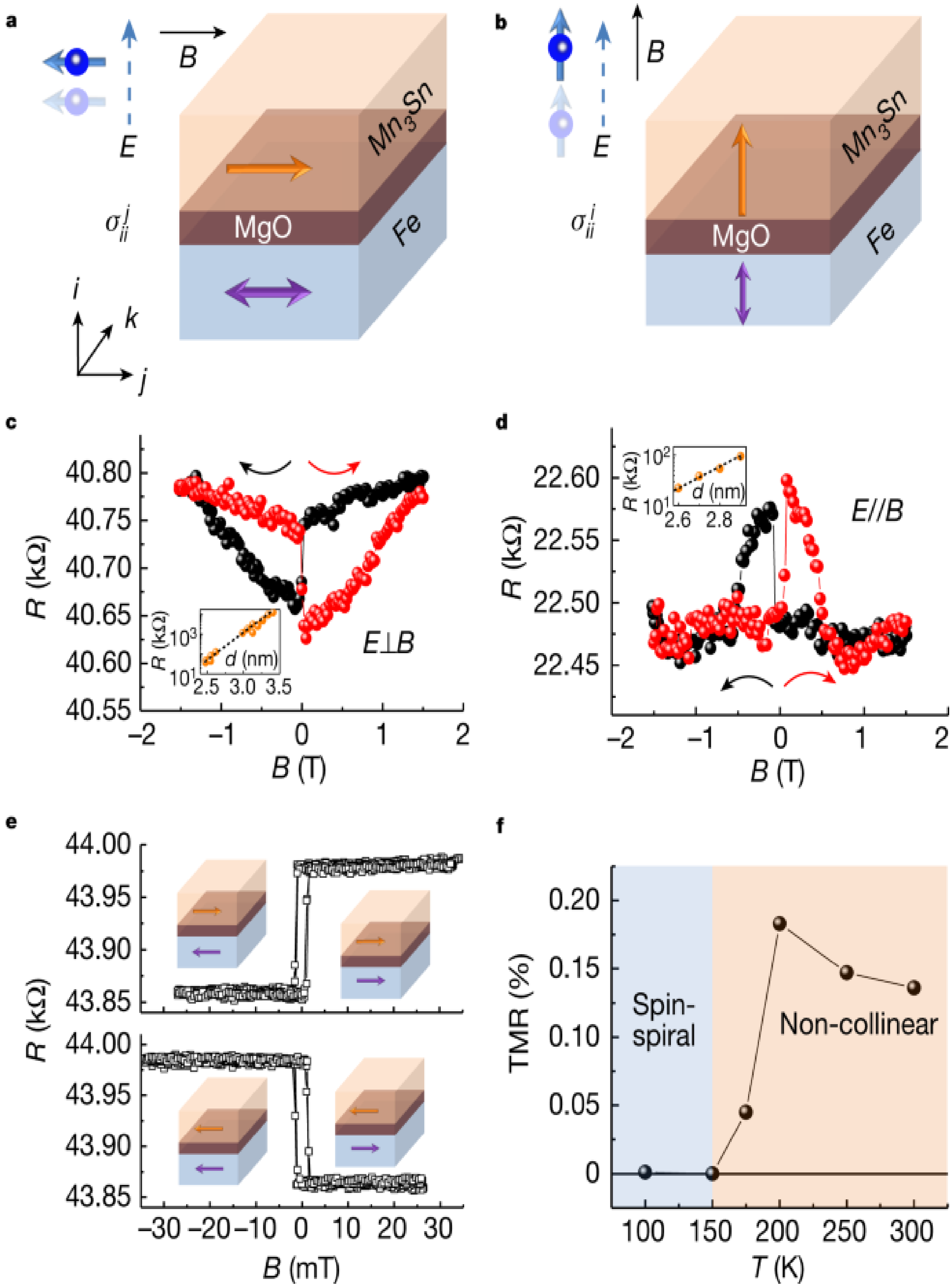
at room temperature in this device. The inset in Fig. 2e shows tunnelling resistance as a function of MgO thickness; the exponential increase in resistance confirmed the reliability of tunnelling conduction.

Our first observation of TMR in the all-antiferromagnetic tunnel junction Mn_3Sn/MgO/Mn_3Sn is outstanding. On the other hand, the magnitude is still small, raising the question of whether this arose from the TRS-breaking octupole polarization—as our theory suggests—or from spin polarization, which is widely accepted to generate the conventional TMR effects[1,2,3,4,5]. Previous theoretical work indicates that Mn_3Sn may host the longitudinal spin-polarized current[7] because the non-collinear antiferromagnetic state breaks the TRS macroscopically[6,7]. In the following, we examine the latter scenario by investigating spin-polarized current in Mn_3Sn through the conventional TMR scheme[1].

Theoretically, spin-polarized current is described by a spin-conductivity tensor ${\sigma}_{ij}^{k}$, in which electric field *E* along axis *j* drives the spin current flowing along direction *i* with spin polarization aligned along the *k* axis[7]. The non-collinear antichiral ordering produces momentum-dependent spin polarization on the Fermi surface, and thus the application of an electric field may drive net longitudinal spin-polarized current with the spin-polarization direction along the octupole polarization (Supplementary Fig. 2). To evaluate the magnitude and sign of the longitudinal spin-polarized current generated in Mn_3Sn, we fabricated a Fe/MgO/Mn_3Sn MTJ with an in-plane magnetized Fe layer using, respectively, the antiferromagnet Mn_3Sn and the typical ferromagnet Fe as the source and detector of the spin-polarized current. We can use the magnetic field to orient spin polarization by directing octupole polarization along the magnetic field. First, for detection of ${\sigma}_{ii}^{j}$ we applied the in-plane field so that the directions of electric field *E* and magnetic field *B* (*i* and *j*, respectively) were perpendicular (Fig. 3a). By sweeping the magnetic field along the in-plane direction *j* from +1.5 to −1.5 T and back to +1.5 T, a negative TMR ratio of 0.2% was observed (Fig. 3c). Assuming the spin polarization of Fe/MgO to be around 0.6, analysis based on the Julliere model yields a spin polarization of Mn_3Sn of about -2×10^{-3} (Methods)[1]. Here, the spin-polarization direction (solid blue arrow in Fig.

3a) is perpendicular to the electric field (solid blue arrow) and the negative sign of spin polarization is consistent with the theoretical analysis of ${\sigma }_{ii}^{j}$ (ref. [7]).

Fig. 3: Anisotropic longitudinal spin-polarized current induced by electric current from Mn_3Sn.



a
B
E
σ_{ii}^{j}
Mn3Sn
MgO
Fe
i
k
j
b
B
E
σ_{ii}^{j}
Mn3Sn
MgO
Fe
c
R (kΩ)
B (T)
E⊥B
d (nm)
d
R (kΩ)
B (T)
E//B
d (nm)
e
R (kΩ)
B (mT)
f
TMR (%)
T (K)
Spin-spiral
Non-collinear

a,**b**, Schematics of the MTJ consisting of Fe/MgO/Mn_3Sn. The spin current detector Fe has the in-plane easy axis (**a**) and out-of-plane easy axis (**b**) (Methods). The corresponding magnetizations in Fe (purple arrows), net polarization direction of octupole polarization (orange arrows) and spin polarization (small blue arrows) of longitudinal spin current induced along the electric field *E* (dashed blue arrows) in Mn_3Sn are schematically shown. The schematics also show switching of magnetization in Fe with a small coercivity of approximately a few mT. **c**, Field dependence of TMR measured in magnetic field *B* applied within the Mn_3Sn film plane, and electric current *I* applied perpendicular to the Mn_3Sn plane using the MTJ shown in **a**. **d**, Field dependence of TMR measured in magnetic field *B* and electric current *I* applied perpendicular to the Mn_3Sn film plane using the MTJ shown in **b**. Red (black) curved arrow indicates the direction of the *B* sweep for obtaining TMR data shown in red (black). Insets show tunnelling resistance *R* as a function of MgO thickness *d*. **e**, Magnetic field dependence of TMR measured using the MTJ shown in **a**. The maximum field is 30 mT so that only the Fe moment switches. Results shown in upper and lower panels were obtained after saturating the octupole polarization in Mn_3Sn with a magnetic field of +1.5 and −1.5 T, respectively. **f**, Temperature dependence of TMR ratio measured using the MTJ shown in **a**. The TMR exists only when Mn_3Sn is in the non-collinear antichiral phase (pink region). This behaviour characteristic of Mn_3Sn indicates that the Mn_3Sn interfacing MgO and leading to the TMR has properties similar to bulk Mn_3Sn (Methods).

Source data

Second, we measured the TMR in another series of Fe/MgO/Mn_3Sn MTJs with a perpendicular magnetized Fe layer as a function of magnetic field *B* along the out-of-plane direction parallel to electric field *E*, for detection of \({\sigma }_{ii}^{i}\) (Methods and Fig. 3b). In contrast with the in-plane geometry above, a positive ratio of 0.5% (Fig. 3d) was observed in the out-of-plane magnetized Fe/MgO/Mn_3Sn. Our analysis found the spin polarization of Mn_3Sn to be approximately $+4 \times 10^{-3}$, consistent with the predicted positive sign of \({\sigma }_{ii}^{i}\) for the case in which the spin polarization direction (small blue arrow in Fig. 3b) is parallel to the

electric field (blue dashed arrow). The bias dependence of TMR (Extended Data Fig. 6), which may give useful information about the energy dependence of polarization[41], clarified the symmetric bias dependence of the TMR ratio for the perpendicular ($E \perp B$) and parallel ($E // B$) configurations (Extended Data Fig. 7), confirming the anisotropy of the longitudinal spin-polarized current in Mn_3Sn.

To further confirm the TRS-breaking character of the longitudinal spin-polarized current, we examined the TMR in the configuration shown in Fig. 3a after polarizion of the octupoles in Mn_3Sn with an in-plane magnetic field, B of +1.5 T and −1.5 T. For each polarized state we measured the TMR by sweeping the in-plane magnetic field within 30 mT in which only Fe moments were switched. The two different polarities of TMR were clearly observed, corresponding to polarization in +1.5 T (Fig. 3e, top) and −1.5 T (Fig. 3e, bottom), in accordance with the TRS-breaking character of the spin current causing the hysteresis shown in Fig. 3c. Moreover, the TMR in Fe/MgO/Mn_3Sn exhibited temperature dependence characteristic of Mn_3Sn (Fig. 3f); it is only in the antichiral non-collinear phase in which the octupoles form a ferroic order that can generate the spin-polarized current[7]. Thus, the observed temperature dependence confirms that the TMR effect in Fe/MgO/Mn_3Sn is driven by the spin-polarized current generated in Mn_3Sn.

Finally, the spin polarization experimentally obtained for Mn_3Sn allowed us to estimate the TMR ratio of Mn_3Sn/MgO/Mn_3Sn based on the Julliere model[1] (Methods). Namely, if we assume that the TMR is based on tunnelling between spin-polarized states as in conventional ferromagnets, the ratio should be about 0.002%, which is more than two orders of magnitude lower than the value of around 0.6% shown in Fig. 2e. In addition, with increased barrier thickness and tunnelling resistance area we found that the ratio was increased up to 1.6% (Extended Data Fig. 8 and Methods). This indicates that the TMR observed in Mn_3Sn/MgO/Mn_3Sn originated not from the spin-polarized current or weak ferromagnetic moments due to canting, but from TRS breaking in the antiferromagnet—that is, the momentum-dependent, spin-splitting bands[7,32,35,36] caused by the polarization of magnetic octupoles[37].

In conclusion, we have demonstrated that the all-antiferromagnetic tunnel junction Mn_3Sn/MgO/Mn_3Sn shows the TMR effect at a ratio of about 2% at room temperature between the parallel and antiparallel alignments of octupole moments. The TMR ratio is far larger than its estimate based on the anisotropic spin polarization found for Mn_3Sn. Besides, our theory predicts that the antiferromagnetic TMR effect originating from the TRS-breaking magnetic octupole may become as large as conventional TMR commonly seen in ferromagnets. Thus, our first observations of the all-antiferromagnetic TMR effect, as well as the anisotropic spin-polarized current, provide new and useful functionality, opening new directions in the research and development of antiferromagnetic spintronics and magnetic memory technology.

Methods

Sample growth

Mn_3Sn/MgO/Mn_3Sn MTJs

A W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm)/Mn_3Sn (42 nm)/Ta (5 nm) (left: substrate side, right: surface side) multilayer was grown on a MgO(001) substrate. The W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm) layer was fabricated by the molecular beam epitaxy (MBE) method under ultrahigh vacuum (UHV) at a base pressure of 2×10^{-8} Pa. The MgO(001) substrate was annealed at 800 °C for 10 min in the MBE chamber before deposition. The W layer (9 nm) was deposited at a rate of 0.1 Å s^{-1} at 300 °C and subsequent annealment at 800 °C for 10 min. The Mn_3Sn layer (12 nm) was fabricated at a rate of 0.25 Å s^{-1} with coevaporation of Mn and Sn, in which the deposition rate of Mn and Sn was set for the stoichiometric composition Mn_3Sn. The Mn_3Sn layer (3 nm) was first deposited at room temperature and then annealed at 320 °C. The extra Mn_3Sn layer (9 nm) was deposited at approximately 260 °C. Subsequently the MgO layer (3.3 nm) was fabricated at a rate of 0.1 Å s^{-1} at room temperature. The stack was later annealed at 600 °C for 30 min. As shown in Extended Data Fig. 1a, streak patterns were observed by reflection high-energy electron diffraction (RHEED),

confirming the formation of epitaxial growth of flat interfaces in the W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm) layer. The incident electron beam is parallel to the MgO [100] direction. The stack was transferred to a magnetron sputtering chamber with base pressure greater than 5×10^{-7} Pa. In the sputtering chamber, the Mn_3Sn (42 nm)/Ta (5 nm) layer was grown at room temperature by magnetron sputtering at a rate of 0.1 nm s^{-1} and a power of 60 W and Ar gas pressure of 0.5 Pa. After deposition, the entire stack was annealed at 450 °C to crystallize the Mn_3Sn layer (42 nm), similar to our previous work for polycrystalline Mn_3Sn[42].

To investigate crystallinity and orientation, cross-sectional transmission electron microscopy (TEM) images for the W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm)/Mn_3Sn (42 nm)/Ta (5 nm) multilayer were taken at room temperature using a commercial TEM system (JEOL, JEM-ARM200F). The maximum operating voltage was 200 kV. Samples for TEM observation were prepared from the TMR device, consisting of the multilayer. using a focused ion beam (Hitachi High-Tech NX2000, Ga 2–30 kV, Ar 1 kV). Before processing, a protective film (C (100 nm)/W (100 nm)) was deposited on an area of 10×3 μm^2 on the sample surface and the sample subsequently thinned by focused Ga and Ar ion beams. The TEM images presented in Fig. 2a and Extended Data Fig. 10a show the sharp interface between the Mn_3Sn layers and the MgO layer. Nanobeam electron diffraction patterns of the Mn_3Sn (top), MgO, Mn_3Sn (bottom) and W layers show epitaxial growth from the W layer to the MgO barrier (Extended Data Fig. 10b–e). As shown in Extended Data Fig. 10a, we fabricated the epitaxial Mn_3Sn layer on the MgO substrate(001)[010]||W$(001)[\bar{1}10]$ stacks that have the $(01\bar{1}1)$ orientation aligned close to the direction of thickness. In this orientation the kagome plane, which is the magnetic easy plane for the cluster magnetic octupole, is oriented along nearly 60° off the normal direction of the film—that is, the *c* axis of Mn_3Sn is about 30° off the normal direction of the film, as shown by the transparent green plane in Extended Data Fig. 10f. Here the size of the $(01\bar{1}1)$-oriented Mn_3Sn crystallite was confirmed to be approximately 100 nm. The MgO barrier on the epitaxial Mn_3Sn (bottom) layer has around 10 nm crystallites and shows the (001) orientation with a mosaicness of about 10°.

An atomic arrangement of each layer shows a potential epitaxial relationship (Extended Data Fig. 10g) in which the lattice constants of MgO, W and Mn_3Sn are considered to be $a = 4.21$ Å for MgO, $a = 3.17$ Å for W[43] and $a = 5.66$ Å and $c = 4.53$ Å for Mn_3Sn[6]. Whereas the Mn arrangement is not perfectly square and may thus have 90°-rotated variants in the film plane, the out-of-plane orientation for all variants should be similar to that observed in the present study (Extended Data Fig. 10a).

Fe/MgO/Mn_3Sn MTJ with an in-plane magnetized Fe layer

A MgO (5 nm)/Fe (30 nm)/MgO (about 3 nm)/Mn_3Sn (42 nm)/Ta (5 nm) multilayer was deposited on the MgO(001) substrate. The MgO (5 nm)/Fe (30 nm)/MgO (about 3 nm) was fabricated using the MBE method under UHV, as described above for the Mn_3Sn/MgO/Mn_3Sn deposition process. The MgO(001) substrate was annealed at 800 °C for 10 min in the MBE chamber before deposition. First, a MgO layer (5 nm) was grown on the substrate at a rate of 0.1 Å s^{-1} at room temperature. Next, the Fe layer (30 nm) was deposited at a rate of 0.25 Å s^{-1} at room temperature and subsequently annealed at 350 °C for 15 min. Finally, the MgO layer (about 3 nm) was fabricated at a rate of 0.1 Å s^{-1} at room temperature. As shown in Extended Data Fig. 1b, clear streak patterns were observed by RHEED, confirming the formation of epitaxial layers with flat interfaces in the MgO (5 nm)/Fe (30 nm)/MgO (about 3 nm) layer. The incident electron beam was parallel to the MgO [100] direction. The stack was transferred to the magnetron sputtering chamber and the Mn_3Sn (42 nm)/Ta (5 nm) layer additionally grown at room temperature by magnetron sputtering at a rate of 0.1 nm s^{-1} with a power of 60 W and Ar gas pressure 0.5 Pa. After deposition, the entire stack was annealed at 450 °C for 30 min.

Fe/MgO/Mn_3Sn MTJ with a perpendicular magnetized Fe layer

A MgO (5 nm)/V (30 nm)/Fe (0.6 nm)/MgO (about 3 nm)/Mn_3Sn (42 nm)/Ta (5 nm) multilayer was deposited on the MgO(001) substrate. The MgO (5 nm)/V (30 nm)/Fe (0.6 nm)/MgO (about 3 nm) was fabricated by the MBE method in the UHV MBE chamber. Similar to the method used for the

$Fe/MgO/Mn_3Sn$ stack, we first deposited a MgO layer (5 nm) on the substrate at a rate of 0.1 Å s^{-1} at room temperature. The V layer (30 nm) was deposited at a rate of 0.25 Å s^{-1} at room temperature and subsequently annealed at 500 °C for 20 min. The Fe (0.6 nm) and MgO (about 3 nm) layers were deposited at room temperature at rates of 0.05 and 0.1 Å s^{-1}, respectively. As shown in Extended Data Fig. 1c, clear streak patterns were observed by RHEED, confirming the formation of epitaxial layers with flat interfaces in the MgO (5 nm)/V (30 nm)/Fe (0.6 nm)/MgO (about 3 nm) layer. The incident electron beam was parallel to the MgO [100] direction. The stack was transferred to the magnetron sputtering chamber and the Mn_3Sn (42 nm)/Ta (5 nm) layer additionally grown by magnetron sputtering at room temperature at a rate of 0.1 nm s^{-1} with a power of 60 W and Ar gas pressure 0.5 Pa. After deposition, the entire stack was annealed at 450 °C for 30 min. In this stack the V layer was used as the seed layer, which can help to induce the strong perpendicular magnetic anisotropy of ultrathin Fe (under 1 nm)[44,45].

We carried out all fabrication processes in situ, including sample transfer from the MBE chamber to the sputtering chamber. The sputtering-grown Mn_3Sn layer (42 nm) on the MgO layer was used for all samples. The composition of this Mn_3Sn layer was determined to be $Mn_{3.15}Sn_{0.85}$ by scanning electron microscopy–energy-dispersive X-ray spectroscopy. Although this composition is Mn rich, it is also a stable composition for a single phase of the $D0_{19}$ Mn_3Sn[42], in which excess Mn randomly occupies the Sn site[6,46].

MTJ fabrication and magnetic and transport measurement

We encapsulated MTJ devices in SiO_x, with electrical contacts formed from 60 nm Pt. Hall measurements were conducted at 300 K in a commercial physical property measurement system (Quantum Design). The field dependence of Hall resistivity was obtained after subtracting the longitudinal resistivity contribution, which was found to be constant as a function of magnetic field. Zero-field Hall resistivity, $\rho_H(B = 0)$ was estimated as $(\rho_H(B = +0) - \rho_H(B = -0)/2$. Here +0 and −0 were used to indicate zero magnetic

field approached from +2 and −2 T, respectively. Tunnelling resistance was measured with a two-probe method in a probe system with an electromagnet at room temperature. A commercial source measurement unit (Keithley 2400, Tektronix) was used to measure resistance for the microfabricated magnetic tunnel junctions. Because a fixed resistor of 1 kΩ was series connected with the MTJ to protect it during measurements, MTJ data include 1 kΩ from the fixed resistor. Electrical measurements recorded below room temperature were performed under 10^{-4} Pa in a vacuum chamber cooled by a helium compressor. The sample substrate was fixed on a Cu sample stage and temperature measured by a thermometer inside the sample stage.

MOKE magnetometry

The magnetic field dependence of MOKE was measured using a commercial system (NanoMOKE3, Quantum Design). The top and bottom Mn_3Sn layers in the W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm)/Mn_3Sn (42 nm)/Ta (5 nm) stacks were used for measurement in a polar MOKE configuration under out-of-plane applied magnetic fields between –1.3 and +1.3 T at room temperature. MOKE loops were acquired using a 660 nm semiconductor laser and a spatial light modulator enabling acquisition of 20 hysteresis loops at a rate of 0.1 Hz. To obtain the MOKE signal, we subtracted the *B*-linear part originating from the extrinsic contribution (for example, a Faraday effect of the optical lenses). Our MOKE measurements for the W (9 nm)/Mn_3Sn (12 nm)/MgO (3.3 nm)/Mn_3Sn (42 nm)/Ta (5 nm) stacks showed that the coercive field of B_c (approximately 0.5 T) for the top Mn_3Sn (Fig. 2c, red) was twice that of B_c (about 0.25 T) for the bottom layer (Fig. 2c, blue), probably due to the difference in buffer layers facing each Mn_3Sn (W for bottom and MgO for top), as well as to variation in the thickness of Mn_3Sn.

Calculation of the projected density of states

The projected density of states (pDOS) of Mn_3Sn and body-centred cubic Fe (bcc-Fe) was calculated with the Wannier functions obtained using the WANNIER90 package[47,48,49], in which localized Wannier functions are

constructed by projection of Bloch wave functions onto atomic orbitals. Bloch wave functions were obtained by density functional theory (DFT) calculations using the QUANTUM ESPRESSO (QE) packages[50,51]. In DFT calculations, the projector-augmented wave pseudopotential with spin-orbit couplings[52] was used, with exchange correlation taken into account by Perdew–Burke–Ernzerhof-type generalized gradient approximation[53]. For Mn_3Sn, the lattice constants $a = 5.665$ Å and $c = 4.531$ Å were used[54] and k-point meshes were $7 \times 7 \times 7$ and $8 \times 8 \times 8$ for self-consistent field (scf) and non-scf calculations, respectively. The cut-off energies of wave function and charge density were 80 and 320 Ry, respectively. Bloch wave functions were projected onto the *s*-, *p*- and *d*-orbitals of Mn ions and the *s*- and *p*-orbitals of Sn ions. For bcc-Fe we used the lattice constant $a = 2.87$ Å, and $8 \times 8 \times 8$ and $12 \times 12 \times 12$ k-point grids were used for scf and non-scf calculations, respectively. We set the energy cut-offs of wave function and charge density as 80 and 500 Ry, respectively. Bloch states were projected onto the *s*-, *p*- and *d*-orbitals of the Fe ion. In Wannierization we set the k-point mesh as $8 \times 8 \times 8$ and $12 \times 12 \times 12$ for Mn_3Sn and bcc-Fe, respectively. Using Wannier functions we calculated pDOS with the $64 \times 64 \times 64$ k-point grid onto cluster magnetic octupolar ordered states of Mn_3Sn and the magnetic ordered states of bcc-Fe.

Cluster magnetic octupole and its polarization

For the symmetry operation of the structural D_{6h} point group, the non-collinear magnetic order of Mn_3Sn has the same transformation properties as the cluster magnetic octupole moment, with ${T}_{\gamma }^{x}=\frac{1}{\sqrt{2}}(-{M}_{31}+{M}_{3-1})$ and ${T}_{\gamma }^{y}=\frac{i}{\sqrt{2}}({M}_{31}+{M}_{3-1})\,$[37,55]. This octupole moment has the same irreducible representation as the ferromagnetic dipole, ${J}_{x}=\frac{1}{\sqrt{2}}(-{M}_{11}+{M}_{1-1})$ and ${J}_{y}=\frac{i}{\sqrt{2}}({M}_{11}+{M}_{1-1})$, and thus its ferroic order breaks time-reversal symmetry macroscopically. This also indicates that the cluster magnetic octupole is parallel to the weak ferromagnetic moment (approximately 7 $m\mu_B$/Mn) due to spin canting, which arises as a result of the competition between Dzyaloshinskii–Moriya interaction, exchange coupling and single-ion anisotropy[56]. Therefore, the driving

mechanism of the anomalous Hall effect of Mn_3Sn can be interpreted as the ferroic order of the octupole moment of Mn_3Sn, in the same way as the ferromagnetic order of the dipole moment of Fe. The cluster multipole theory is thus useful for understanding the underlying physics on non-collinear antiferromagnets such as Mn_3Sn[37]. In the calculation of octupole polarization, we estimated the expectation value of the following operator with $p = 3$ and $q = \pm 1$ for the Bloch wave functions obtained by generalized gradient approximation calculation:

$${\tau }_{pq}^{(\mu)}\equiv \sqrt{\frac{4\pi }{2p+1}}\mathop{\sum }\limits_{i=1}^{{N}_{{\rm{atom}}}^{(\mu)}}{{\boldsymbol{\sigma }}}_{i}\cdot {\nabla }_{i}(|\,{{\bf{R}}}_{i}\,{|}^{p}\,{Y}_{pq}{({\theta }_{i},{\varphi }_{i})}^{* })$$
(1)

in which ${N}_{{\rm{atom}}}^{(\mu)}$ is the number of atoms of the μth cluster, $\boldsymbol{\sigma}_i$ is the Pauli matrices defined for the spin degrees of freedom of the ith atom, ${\nabla }_{i}\equiv \frac{\partial }{\partial {{\bf{R}}}_{i}}$, $\mathbf{R}_i \equiv (X_i, Y_i, Z_i)$ is the position of the ith atom, Y_{pq} are the spherical harmonics and $\mathbf{R}_i$, θ_i and ϕ_i are the distance, polar angle and azimuthal angle, respectively, of the ith atom.

Estimation of spin polarization of Fe and Mn_3Sn using the Julliere model

Based on the Julliere model[1], the relative conduction change, ΔG between states 0 and 1 is described as $\Delta G = G \times 2P_1P_2/(1 + P_1P_2)$, where G, P_1 and P_2 correspond, respectively to the tunnelling conduction and spin-polarization ratio of the effective tunnelling density of states of magnetic electrodes 1 and 2. The spin polarization of Fe/MgO in our device is assumed to be about 0.6 (ref. [45]).

Fermi level shift and resultant negative TMR

Given the strong Fermi energy dependence on the sign of octupole polarization (Extended Data Fig. 2c), the sign of TMR should be very sensitive to the Fermi energy of Mn_3Sn. Therefore, interfacial engineering may influence the Fermi energy of Mn_3Sn. Extended Data Fig. 4a shows the Hall resistivity of Mn_3Sn as a function of temperature. When decreasing the temperature, Hall resistivity vanishes due to the phase transition from non-collinear antichiral to spin-spiral. Both the anomalous Hall effect and non-zero spin polarization exist only in the non-collinear antichiral phase of Mn_3Sn rather than in the spin-spiral phase. Nevertheless, phase transition temperature detected through TMR (black circle) was around 100 K lower than that (blue square) found in Hall resistivity (Extended Data Fig. 4a). This is most probably because the Fermi energy of the interfacial Mn_3Sn shifted as compared with that of the Mn_3Sn film, due to its contact with MgO. In addition, the distinct Fermi level shift of Mn_3Sn at the bottom and top may have arisen due to variation in the thermal annealing process (Methods). Here we roughly simulated normalized TMR via octupole polarizations from bottom τ_{bottom} and top Mn_3Sn τ_{top}, normalized TMR = $\tau_{\text{bottom}} \times \tau_{\text{top}}$, assuming that the difference in Fermi energy shift is 1 eV (Extended Data Fig. 4b).

First-principles calculation of the tunnelling magnetoresistance effect with Mn3Sn electrodes

To theoretically simulate the TMR effect from first principles, we use the PWCOND package in the QE package[57,58,59] in which ballistic transport along the *z* direction is calculated by solving the scattering problem on the Bloch wave function obtained by DFT calculation[60]. Generally for TMR, it is important to have in-depth understanding of the barrier material[61,62,63]. On the other hand, before investigating the role of the barrier material it is crucial to determine whether there is any TMR effect in the all-antiferromagnetic tunnel junction. Thus, in our calculation, for simplicity we use Mn_3Sn for the electrodes and the vacuum for the barrier. In practice, we calculate the tunnelling conductance in the Mn_3Sn/vacuum/Mn_3Sn MTJ. The MTJ in our calculations is made by stacking Mn_3Sn along the *c* axis: the

conducting path is perpendicular to the *ab* plane of Mn_3Sn. The schematics of the MTJ are shown in Extended Data Fig. 3a. We calculated transmissions for both parallel and antiparallel configurations; the cluster magnetic octupole moments of the two electrodes point in the same and opposite directions in the parallel and antiparallel configurations, respectively.

First, we separate the entire MTJ system into three parts: left and the right leads consisting of the bulk Mn_3Sn and the scattering region, comprising a pair of two monolayers of Mn_3Sn and the vacuum region between. We performed the DFT calculation with QE for each of the three parts. We set the *k*-point mesh as $7 \times 7 \times 7$ for the leads and $7 \times 7 \times 1$ for the scattering region. In calculation of the scattering region, the constraint on magnetic moments was imposed to stabilize the magnetic structure. When we calculated the electronic structure of the scattering region in the antiparallel configuration, to smoothly connect the leads and the scattering region in the transmission calculation we treated the doubled scattering region, which consists of the original scattering region and its copy, attached to the original one with its magnetic configuration inverted. In this way we calculated the electronic structure of the Mn_3Sn/vacuum/Mn_3Sn/Mn_3Sn/vacuum/Mn_3Sn system. The doubled scattering region was cut in half and only the original was considered in transmission calculations.

Then, connecting the leads and scattering region, we calculated transmission. We set the $k_\perp = (k_x, k_y)$ point mesh in the *xy* plane as 32×32. We obtained transmissions at each $k_\perp$-point for the parallel/antiparallel configurations, $T_{\rm P/AP}(k_\perp)$; conductances for parallel/antiparallel configurations, $G_{\rm P/AP}$, were determined by the Landauer–Büttiker formula[64,65,66,67], ${G}_{{\rm{P/AP}}}=({e}^{2}/h){\Sigma }_{{{\boldsymbol{k}}}_{\perp }}{T}_{{\rm{P/AP}}}({{\boldsymbol{k}}}_{\perp })$. The TMR ratio was calculated as $(G_{\rm P} - G_{\rm AP})/G_{\rm AP}$.

The vacuum thickness, *d*, dependence of the total transmission at the Fermi level for the parallel and antiparallel configurations, is shown in Extended Data Fig. 3b. We also plotted the resistance-area product (RA) for each configuration, which is the normalized resistance given as $({\rm RA}) = A/G$,

where A is the cross-section area. We found that G_P was larger than G_{AP} in all cases of d, and both G_P and G_{AP} almost exponentially decayed with d, as shown in Extended Data Fig. 3b,c. We thus obtained the positive TMR ratio (Extended Data Fig. 3d). Transmissions resolved by $k_\perp$-points are shown in Extended Data Fig. 3e,f, which indicates that transmission behaviour does differ between the parallel and antiparallel configurations.

For comparison with the Mn_3Sn/vacuum/Mn_3Sn MTJ, we also investigated the transmission properties of a ferromagnetic MTJ with the vacuum barrier, the Fe/vacuum/Fe system. In the same manner as for the Mn_3Sn/vacuum/Mn_3Sn MTJ, we calculated electronic structures without spin-orbit couplings of the leads composed by the bulk bcc-Fe with $8 \times 8 \times 8$ k-mesh, and of the scattering region which has the vacuum sandwiched between a pair of four monolayers of Fe with $8 \times 8 \times 1$ k-mesh. We performed the transmission calculation using 100×100 $k_\perp$-point grids. Extended Data Fig. 3g,h shows the d-dependence of total transmissions and RA values at the Fermi level for both parallel and antiparallel arrangements. The TMR ratio with respect to d is plotted in Extended Data Fig. 3i, taking as large a value as that in the Mn_3Sn/vacuum/Mn_3Sn MTJ. The results should be sufficient to serve as a qualitative reference for the Mn_3Sn/vacuum/Mn_3Sn MTJ, whereas a small non-monotonic change in the TMR ratio was observed, which could converge by the calculation with higher accuracy.

The electronic states that dominate tunnelling transport correspond to electrons tunnelling in the normal direction to the interface, and thus the polarization of such states is important for discussion of tunnelling physics[68]. In fact, the tunnelling conductance of the Fe/MgO/Fe MTJ has the peak at $k_\perp$ approximately 0, supporting this concept[61,62]. In a more complex system such as Mn_3Sn, however, the states whose group velocity carries only normal incidence components exist not only at $k_\perp$ approximately 0 but also at general k points[69], and such states should largely contribute to tunnelling conductance following the concepts proposed by Slonczewski[68]. Extended Data Fig. 3e,f shows the in-plane momentum (k_x, k_y) dependence of transmission integrated over momentum along the tunnelling direction.

Notably, our results clarify that transmission involves not only those states at $k_\perp$ approximately 0, but also of the widely extended momentum region in the Brillouin zone. Given that the results in the figures are those projected to the in-plane momentum, tunnelling electrons arise not only from momentum $k_\perp$ of approximately 0 related to normal incidence, but rather from entire region of the Brillouin zone. Thus, to qualitatively understand the mechanism of tunnelling conductance, we should focus not only on the states with $k_\perp$ approximately 0 but rather use the measure reflecting contributions from the entire Brillouin zone. Given the fact that the antiferromagnetic state of Mn_3Sn can be viewed as the ferroic order of cluster magnetic octupole, such a measure should be the summation of octupole polarization over the entire momentum space—the density of states projected onto octupole polarization.

Effect of interfacial structures on transmission properties

We investigated the robustness of transmission properties against variation in interfacial structures in the Mn_3Sn/vacuum/Mn_3Sn MTJ. We examined two types of variation of the interface: interfacial disorder and lateral shift.

First we studied the effect of disorders. We incorporated interfacial disorders into the Mn_3Sn/vacuum/Mn_3Sn MTJ by artificially moving some atoms facing the vacuum barriers; we shifted upwards one of the Mn atoms in the lower layer of Mn_3Sn at the interface, and shifted downwards one of the Mn atoms in the upper layer of Mn_3Sn. These atoms were shifted by 0.113 Å—that is, 2.5% of the *c* axis length of Mn_3Sn, which does not qualitatively change the electronic and magnetic properties of the system. Extended Data Fig. 5a–c shows the results of calculations against the disorder with barrier thickness 4.531 Å. Whereas shifting of atoms decreased transmission in the parallel configuration and increased it in the antiparallel configuration, overall transmission properties did not change. We also examined the case in which atoms move inversely to the case above: the atom in the lower layer moved downwards whereas that in the upper layer moved upwards. The results of these calculations (Extended Data Fig. 5d–f) suggest that transmission properties did not qualitatively change in this case either. Whereas it is better to use a larger supercell for more precise evaluation of

interfacial disorder, we expect that TMR properties may not largely change from those shown here.

Second, we examined the effect of lateral shift. Mn_3Sn has two layers in a unit cell along the *c* axis—say, A and B. In the calculations whose results are shown in Fig. 3, layers A and B are at the interface, which we call geometry-I. Here we also consider the case in which two B-layers face each other (see the inset in Extended Data Fig. 5i), which we call geometry-II. We show the results of the calculations with geometries-I and II in Fig. 5g–i. These results indicate that the properties in tunnelling conductance do not largely change, whereas total transmissions in geometry-II take larger values than those in geometry-I. We note that for geometry-II we take the doubled unit cell both for the parallel and antiparallel configurations, due to its geometry. We confirmed that the finite TMR effect is observed for other lattice-matching configurations—that is, when two electrodes have differently oriented easy axes of cluster magnetic octupoles by 120° in geometry-I (Supplementary Fig. 4). We also examined the TMR effect by shifting the upper layer of Mn_3Sn along the *a* axis by half of the lattice constant *a*, and confirmed that this shift qualitatively maintained the TMR effect (Supplementary Fig. 5).

Bias-dependent TMR measurements

The bias dependence of TMR can provide useful information about the energy dependence of polarization[41]. Thus, we performed TMR measurements as a function of bias voltage in the Fe/MgO/Mn_3Sn tunnel junction for both the perpendicular ($E \perp B$) and parallel (E//B) configurations for the electric (E) and magnetic fields (B), corresponding to Fig. 3c,d, respectively. In the following, we show the experimental summary for the perpendicular ($E \perp B$) case as a representative example (Extended Data Fig. 6). Extended Data Fig. 6a,b shows the minor loops under bias voltage of +0.6 and −0.5 V, in which Fe moments are switched while Mn_3Sn moments remain fixed.

The key observation is the symmetric bias dependence of the TMR ratio on the perpendicular and parallel configurations (Extended Data Fig. 7). This becomes clear if we consider a bias-independent term of the order of around 0.14 % (horizontal dashed line). According to our measurement

configuration, the hot electron arises from Mn_3Sn and thus MR should reflect the unoccupied DOS of Fe for positive bias (Extended Data Fig. 7a). In this regime, the spin polarization of Mn_3Sn should be determined by the DOS of Mn_3Sn in the vicinity of the Fermi level. Spin polarization around the Fermi level in Mn_3Sn is anisotropic and can be characterized by the tensor, as discussed above. Thus, MR should have negative and positive signs for perpendicular and parallel configurations, respectively (Extended Data Fig. 7b). This explains the following observations of our experiment. Namely, after subtracting the constant background MR term of around 0.14%, we find that MR has magnitudes similar to the opposite signs for perpendicular and parallel configurations. Bias dependence must arise from the unoccupied DOS of Fe and should be similar to that observed in the Fe/MgO/Fe MTJ[5].

On the other hand, negative bias should provide the hot electron from Fe to Mn_3Sn. Thus, MR would be determined by the spin polarizations of unoccupied DOS for Mn_3Sn and DOS for Fe in the vicinity of the Fermi level. As compared with the positive bias regime, MR in the negative bias is strongly bias dependent (Extended Data Fig. 7b). This would be due to the bias voltage dependence of the anisotropic spin polarization in Mn_3Sn. Interestingly, the MRs for both perpendicular and parallel MTJs converge at around 0.1% at bias V below −0.5. Most probably, this is a result of isotropic spin polarization of Mn_3Sn from spin canting.

Measurement of the thickness dependence of TMR

The TMR in Mn_3Sn/MgO/Mn_3Sn MTJ was measured by variation in MgO thickness. When the barrier became thicker, TMR value increased, consistent with the theoretical calculations shown in Extended Data Fig. 2. We also performed a comparison between experiments and calculations, as shown in Extended Data Fig. 8. Whereas these calculations use the vacuum barrier, we set the *x* axis to be the resistance area rather than MgO thickness itself, for clarity.

Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

References

1. Julliere, M. Tunneling between ferromagnetic films. *Phys. Lett. A* **54**, 225–226 (1975).

2. Miyazaki, T. & Tezuka, N. Giant magnetic tunneling effect in $Fe/Al_2O_3/Fe$ junction. *J. Magn. Magn. Mater.* **139**, L231–L234 (1995).

3. Moodera, J. S., Kinder, L. R., Wong, T. M. & Meservey, R. Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions. *Phys. Rev. Lett.* **74**, 3273–3276 (1995).

4. Parkin, S. S. et al. Giant tunnelling magnetoresistance at room temperature with MgO (100) tunnel barriers. *Nat. Mater.* **3**, 862–867 (2004).

5. Yuasa, S., Nagahama, T., Fukushima, A., Suzuki, Y. & Ando, K. Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. *Nat. Mater.* **3**, 868–871 (2004).

6. Nakatsuji, S., Kiyohara, N. & Higo, T. Large anomalous Hall effect in a non-collinear antiferromagnet at room temperature. *Nature* **527**, 212–215 (2015).

7. Železný, J., Zhang, Y., Felser, C. & Yan, B. Spin-polarized current in noncollinear antiferromagnets. *Phys. Rev. Lett.* **119**, 187204 (2017).

8. MacDonald, A. & Tsoi, M. Antiferromagnetic metal spintronics. *Philos. Trans. A Math. Phys. Eng. Sci.* **369**, 3098–3114 (2011).

9. Jungwirth, T., Marti, X., Wadley, P. & Wunderlich, J. Antiferromagnetic spintronics. *Nat. Nanotechnol.* **11**, 231–241 (2016).

10. Baltz, V. et al. Antiferromagnetic spintronics. *Rev. Mod. Phys.* **90**, 015005 (2018).

11. Baibich, M. N. et al. Giant magnetoresistance of (001) Fe/(001) Cr magnetic superlattices. *Phys. Rev. Lett.* **61**, 2472–2475 (1988).

12. Chappert, C., Fert, A. & Van Dau, F. N. The emergence of spin electronics in data storage. *Nat. Mater.* **6**, 813–823 (2007).

13. Katine, J. & Fullerton, E. E. Device implications of spin-transfer torques. *J. Magn. Magn. Mater.* **320**, 1217–1226 (2008).

14. Dieny, B. et al. Opportunities and challenges for spintronics in the microelectronics industry. *Nat. Electron.* **3**, 446–459 (2020).

15. Núñez, A. S., Duine, R. A., Haney, P. & MacDonald, A. H. Theory of spin torques and giant magnetoresistance in antiferromagnetic metals. *Phys. Rev. B* **73**, 214426 (2006).

16. Tsai, H. et al. Electrical manipulation of a topological antiferromagnetic state. *Nature* **580**, 608–613 (2020).

17. Takeuchi, Y. et al. Chiral-spin rotation of non-collinear antiferromagnet by spin–orbit torque. *Nat. Mater.* **20**, 1367–1370 (2021).

18. Wadley, P. et al. Electrical switching of an antiferromagnet. *Science* **351**, 587–590 (2016).

19. Chen, X. et al. Antidamping-torque-induced switching in biaxial antiferromagnetic insulators. *Phys. Rev. Lett.* **120**, 207204 (2018).

20. Miron, I. M. et al. Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection. *Nature* **476**, 189–193 (2011).

21. Liu, L. et al. Spin-torque switching with the giant spin Hall effect of tantalum. *Science* **336**, 555–558 (2012).

22. Manchon, A. et al. Current-induced spin-orbit torques in ferromagnetic and antiferromagnetic systems. *Rev. Mod. Phys.* **91**, 035004 (2019).

23. Chen, H., Niu, Q. & MacDonald, A. H. Anomalous Hall effect arising from noncollinear antiferromagnetism. *Phys. Rev. Lett.* **112**, 017205 (2014).

24. Kiyohara, N., Tomita, T. & Nakatsuji, S. Giant anomalous Hall effect in the chiral antiferromagnet Mn_3Ge. *Phys. Rev. Appl.* **5**, 064009 (2016).

25. Nayak, A. K. et al. Large anomalous Hall effect driven by a nonvanishing Berry curvature in the noncolinear antiferromagnet Mn_3Ge. *Sci. Adv.* **2**, e1501870 (2016).

26. Liu, Z. et al. Electrical switching of the topological anomalous Hall effect in a non-collinear antiferromagnet above room temperature. *Nat. Electron.* **1**, 172–177 (2018).

27. Marti, X. et al. Room-temperature antiferromagnetic memory resistor. *Nat. Mater.* **13**, 367–374 (2014).

28. Park, B. G. et al. A spin-valve-like magnetoresistance of an antiferromagnet-based tunnel junction. *Nat. Mater.* **10**, 347–351 (2011).

29. Wang, Y. et al. Room-temperature perpendicular exchange coupling and tunneling anisotropic magnetoresistance in an antiferromagnet-based tunnel junction. *Phys. Rev. Lett.* **109**, 137201 (2012).

30. Chen, X. et al. Tunneling anisotropic magnetoresistance driven by magnetic phase transition. *Nat. Commun.* **8**, 449 (2017).

31. Šmejkal, L., González-Hernández, R., Jungwirth, T. & Sinova, J. Crystal time-reversal symmetry breaking and spontaneous Hall effect in collinear antiferromagnets. *Sci. Adv.* **6**, eaaz8809 (2020).

32. Reichlová, H. et al. Macroscopic time reversal symmetry breaking by staggered spin-momentum interaction. Preprint at https://arxiv.org/abs/2012.15651 (2020).

33. Yang, H. et al. Topological Weyl semimetals in the chiral antiferromagnetic materials Mn_3Ge and Mn_3Sn. *New J. Phys,* **19**,

015008 (2017).

34. Kuroda, K. et al. Evidence for magnetic Weyl fermions in a correlated metal. *Nat. Mater.* **16**, 1090–1095 (2017).

35. Shao, D.-F., Zhang, S.-H., Li, M., Eom, C.-B. & Tsymbal, E. Y. Spin-neutral currents for spintronics. *Nat. Commun.* **12**, 7061 (2021).

36. Šmejkal, L., Hellenes, A. B., González-Hernández, R., Sinova, J. & Jungwirth, T. Giant and tunneling magnetoresistance in unconventional collinear antiferromagnets with nonrelativistic spin-momentum coupling. *Phys. Rev. X* **12**, 011028 (2022).

37. Suzuki, M.-T., Koretsune, T., Ochi, M. & Arita, R. Cluster multipole theory for anomalous Hall effect in antiferromagnets. *Phys. Rev. B* **95**, 094406 (2017).

38. Ikhlas, M. et al. Large anomalous Nernst effect at room temperature in a chiral antiferromagnet. *Nat. Phys.* **13**, 1085–1090 (2017).

39. Higo, T. et al. Large magneto-optical Kerr effect and imaging of magnetic octupole domains in an antiferromagnetic metal. *Nat. Photon.* **12**, 73–78 (2018).

40. Kimata, M. et al. Magnetic and magnetic inverse spin Hall effects in a non-collinear antiferromagnet. *Nature* **565**, 627–630 (2019).

41. Valenzuela, S. O., Monsma, D. J., Marcus, C. M., Narayanamurti, V. & Tinkham, M. Spin polarized tunneling at finite bias. *Phys. Rev. Lett.* **94**, 196601 (2005).

42. Higo, T. et al. Omnidirectional control of large electrical output in a topological antiferromagnet. *Adv. Funct. Mater.* **31**, 2008971 (2021).

43. Featherston, F. H. & Neighbours, J. R. Elastic constants of tantalum, tungsten, and molybdenum. *Phys. Rev.* **130**, 1324–1333 (1963).

44. Lambert, C.-H. et al. Quantifying perpendicular magnetic anisotropy at the Fe-MgO(001) interface. *Appl. Phys. Lett.* **102**, 122410 (2013).

45. Tanaka, K. et al. Large voltage-induced magnetic anisotropy field change in ferrimagnetic FeGd. *Appl. Phys. Express* **8**, 073007 (2015).

46. Brown, P. J., Nunez, V., Tasset, F., Forsyth, J. B. & Radhakrishna, P. Determination of the magnetic structure of Mn_3Sn using generalized neutron polarization analysis. *J. Phys. Condens. Matter.* **2**, 9409–9422 (1990).

47. Mostofi, A. A. et al. wannier90: a tool for obtaining maximally-localised Wannier functions. *Comput. Phys. Commun.* **178**, 685–699 (2008).

48. Mostofi, A. A. et al. An updated version of wannier90: a tool for obtaining maximally-localised Wannier functions. *Comput. Phys. Commun.* **185**, 2309–2310 (2014).

49. Pizzi, G. et al. Wannier90 as a community code: new features and applications. *J. Phys. Condens. Matter* **32**, 165902 (2020).

50. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).

51. Giannozzi, P. et al. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys. Condens. Matter* **29**, 465901 (2017).

52. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

53. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

54. Tomiyoshi, S. Polarized neutron diffraction study of the spin structure of Mn_3Sn. *J. Phys. Soc. Jpn* **51**, 803–810 (1982).

55. Tomiyoshi, S. & Yamaguchi, Y. Magnetic structure and weak ferromagnetism of Mn_3Sn studied by polarized neutron diffraction. *J. Phys. Soc. Jpn* **51**, 2478–2486 (1982).

56. Liu, J. & Balents, L. Anomalous Hall effect and topological defects in antiferromagnetic Weyl semimetals: Mn_3Sn/Ge. *Phys. Rev. Lett.* **119**, 087202 (2017).

57. Smogunov, A., Dal Corso, A. & Tosatti, E. Ballistic conductance of magnetic Co and Ni nanowires with ultrasoft pseudopotentials. *Phys. Rev. B* **70**, 045417 (2004).

58. Corso, A. D. & Conte, A. M. Spin-orbit coupling with ultrasoft pseudopotentials: application to Au and Pt. *Phys. Rev. B* **71**, 115106 (2005).

59. Dal Corso, A., Smogunov, A. & Tosatti, E. *Ab initio* ballistic conductance with spin-orbit coupling: application to monoatomic wires. *Phys. Rev. B* **74**, 045429 (2006).

60. Joon Choi, H. & Ihm, J. Ab initio pseudopotential method for the calculation of conductance in quantum wires. *Phys. Rev. B* **59**, 2267–2275 (1999).

61. Butler, W. H., Zhang, X.-G., Schulthess, T. C. & MacLaren, J. M. Spin-dependent tunneling conductance of Fe|MgO|Fe sandwiches. *Phys. Rev. B* **63**, 054416 (2001).

62. Mathon, J. & Umerski, A. Theory of tunneling magnetoresistance of an epitaxial Fe|MgO|Fe(001) junction. *Phys. Rev. B* **63**, 220403 (2001).

63. Zhang, X. & Butler, W. Theory of giant magnetoresistance and tunneling magnetoresistance. In *Handbook of Spintronics* (eds Xu, Y. et al.) 3–69 (Springer, 2016).

64. Landauer, R. Spatial variation of currents and fields due to localized scatterers in metallic conduction. *IBM J. Res. Dev.* **1**, 223–231 (1957).

65. Landauer, R. Electrical resistance of disordered one-dimensional lattices. *Phil. Mag.* **21**, 863–867 (1970).

66. Büttiker, M. Four-terminal phase-coherent conductance. *Phys. Rev. Lett.* **57**, 1761–1764 (1986).

67. Büttiker, M. Symmetry of electrical conduction. *IBM J. Res. Dev.* **32**, 317–334 (1988).

68. Slonczewski, J. C. Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier. *Phys. Rev. B* **39**, 6995–7002 (1989).

69. Tsymbal, E. Y. et al. Interface effects in spin-dependent tunneling. *Prog. Mater. Sci.* **52**, 401–420 (2007).

Acknowledgements

This work was partially supported by the JST-Mirai Program (no. JPMJMI20A1), JST-CREST Program (no. JPMJCR18T3), JST-PRESTO Program (no. JPMJPR20L7) and JSPS KAKENHI (nos. 21H04437 and 22H00290). X.C. acknowledges support from the Japan Society for the Promotion of Science (Postdoctoral Fellowship for Research in Japan (Standard)). Use of the facilities of the Materials Design and Characterization Laboratory at the Institute for Solid State Physics, the University of Tokyo, is gratefully acknowledged.

Author information

Author notes

1. These authors contributed equally: Xianzhe Chen, Tomoya Higo and Katsuhiro Tanaka

Authors and Affiliations

1. Department of Physics, University of Tokyo, Tokyo, Japan

   Xianzhe Chen, Tomoya Higo, Hanshen Tsai, Hiroshi Idzuchi, Takumi Matsuo & Satoru Nakatsuji

2. Institute for Solid State Physics, University of Tokyo, Chiba, Japan

   Xianzhe Chen, Tomoya Higo, Hanshen Tsai, Hiroshi Idzuchi, Masanobu Shiga, Shoya Sakamoto, Ryoya Ando, Hidetoshi Kosaki, Daisuke Nishio-Hamane, Shinji Miwa & Satoru Nakatsuji

3. CREST, Japan Science and Technology Agency, Saitama, Japan

   Tomoya Higo, Ryotaro Arita, Shinji Miwa & Satoru Nakatsuji

4. Research Center for Advanced Science and Technology, University of Tokyo, Tokyo, Japan

   Katsuhiro Tanaka, Takuya Nomoto & Ryotaro Arita

5. PRESTO, Japan Science and Technology Agency, Saitama, Japan

   Takuya Nomoto

6. RIKEN, Center for Emergent Matter Science, Saitama, Japan

   Ryotaro Arita

7. Trans-scale Quantum Science Institute, University of Tokyo, Tokyo, Japan

   Shinji Miwa & Satoru Nakatsuji

8. Canadian Institute for Advanced Research, Toronto, Ontario, Canada

   Satoru Nakatsuji

Contributions

S.N. conceived and supervised this study. T.H., X.C., H.I., M.S., S.S. and H.K. grew the thin films. X.C. fabricated the devices. X.C. and H.T. carried out magnetotransport measurements. S.N., K.T., T.N. and R. Arita proposed the theoretical calculations. S.N., T.H., H.T., R. Ando, T.M. and S.M. provided suggestions on the experiments. X.C., K.T., T.H. and S.N. wrote

the manuscript together with inputs from T.N., R. Arita and S.M. All authors discussed the results and commented on the manuscript.

Corresponding author

Correspondence to Satoru Nakatsuji.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Guohan Hu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 RHEED images in the deposition process of the multilayers.

a, In-situ reflection high energy electron diffraction (RHEED) images of the MgO (001) substrate after annealing at 800 °C for 10 minutes, the W layer (9 nm) annealed at 800 °C for 10 minutes, the Mn_3Sn layer (total 12 nm) deposited at 260 °C, and the MgO (3.3 nm) layer deposited at room

temperature in the W(9 nm)/Mn_3Sn(12 nm)/MgO(3.3 nm)/Mn_3Sn(42 nm)/Ta(5 nm) multilayer. **b**, RHEED images of the MgO (001) substrate after annealing at 800 °C for 10 minutes, the Fe layer (30 nm) annealed at 350 °C for 15 minutes, and the MgO (~ 3 nm) layer deposited at room temperature in the MgO(5 nm)/Fe(30 nm)/MgO(~ 3 nm)/Mn_3Sn(42 nm)/Ta(5 nm) multilayer. **c**, RHEED images of the MgO (001) substrate after annealing at 800 °C for 10 minutes, the V layer (30 nm) annealed at 500 °C for 20 minutes, the Fe layer (0.6 nm) deposited at room temperature, and the MgO (~ 3 nm) layer deposited at room temperature in the MgO(5 nm)/V(30 nm)/Fe(0.6 nm)/MgO(~ 3 nm)/Mn_3Sn(42 nm)/Ta(5 nm) multilayer. The incident electron beam was parallel to the MgO [100] direction for all the measurements.

Extended Data Fig. 2 Calculation results of the spin polarization of Fe and the octupole polarization of Mn_3Sn.

a, Calculated projected Density of States (pDOS) of Fe for the majority ($pDOS_{maj}$, black) and the minority ($pDOS_{min}$, red) bands with different spin polarizations. **b**, Calculated spin polarization of Fe as a function of the Fermi energy. The spin polarization is normalized by estimating $\sigma = (pDOS_{maj} - pDOS_{min})/(pDOS_{maj} + pDOS_{min})$. **c**, Calculated octupole polarization of Mn_3Sn as a function of the Fermi energy. The normalized octupole polarization is calculated by $\sigma = (pDOS_{maj} - pDOS_{min})/(pDOS_{maj} + pDOS_{min})$. The polarity of octupole polarization changes the sign when $E - E_F$ is between − 1 and 0 eV.

Extended Data Fig. 3 First-principles calculations of the tunneling magnetoresistance (TMR) effect with the vacuum barrier at the Fermi level.

a, Schematics of the Mn_3Sn/vacuum/Mn_3Sn magnetic tunnel junction (MTJ). Blue and red arrows represent the magnetic moments carried by the Mn ions. **b**–**f**, Results for the Mn_3Sn/vacuum/Mn_3Sn MTJ. (**b**), (**c**) Vacuum thickness, *d*, dependence of the (**b**) total transmissions and (**c**) resistance-

area products (RA) in the parallel and antiparallel configurations. (**d**) TMR ratio as a function of d. (**e**), (**f**) Transmissions resolved by the $\boldsymbol{k}_{\perp}$-points for the (**e**) parallel and (**f**) antiparallel configurations at $d = 9.062$ Å, where the hexagons drawn by the black lines denote the first Brillouin zone. (**g**)–(**i**) Results for the Fe/vacuum/Fe MTJ. (**g**) Total transmissions and (**h**) RA values in the MTJ with the parallel and antiparallel configurations, and (**i**) TMR ratio with respect to d.

Extended Data Fig. 4 Interfacial Fermi level shift of Mn_3Sn and resultant negative tunneling magnetoresistance.

a, Evidence for interfacial Fermi level shift of interfacial Mn_3Sn. The phase transition temperature from spin-spiral to noncollinear antichiral states of interfacial Mn_3Sn (indicated by the TMR values (black circle, left)) is lower than the Mn_3Sn films (indicated by the Hall resistivity ρ_{xy} (blue square, right)). **b**, Normalized TMR estimated as TMR = $\tau_{\text{bottom}} \times \tau_{\text{top}}$, assuming that octupole polarizations from bottom is τ_{bottom} and the one from top Mn_3Sn is τ_{top}, and there is 1 eV Fermi level shift between the top and bottom interfacial Mn_3Sn. $E - E_{\text{F}}$ is is the difference between band energy and the Fermi level.

Extended Data Fig. 5 Effects of the interfacial disorder and the lateral shift on the tunneling magnetoresistance (TMR) of the Mn_3Sn/vacuum/Mn_3Sn magnetic tunnel junction (MTJ).

The thickness of the barrier is 4.531 Å. **a,b,c**, Results for the MTJ with the atoms at the lower layer shifted upward and the upper layer downward. The pair of atoms shifted in the lower and upper layers is denoted as (i, j), where i, j = 1, 2, 3 correspond to the Mn-sites shown schematically in the inset in (c). (a) Total transmissions and (b) resistance-area products (RA) in the parallel and antiparallel configurations, and (c) the TMR ratio. **d,e,f**, Results for the MTJ with the atoms at the lower layer shifted downward and the upper layer upward. (d) Total transmissions and (e) RA in the parallel and antiparallel configurations, and (f) the TMR ratio. **g,h,i**, Results for the MTJ with two different geometries at the interface, which is schematically shown

in the inset in (i) where the rhombuses indicate Mn_3Sn. (g) Total transmissions and (h) RA in the parallel and antiparallel configurations, and (i) the TMR ratio.

Extended Data Fig. 6 Bias dependent tunneling magnetoresistance of the Fe/MgO/Mn_3Sn junction.

a,b, Magnetic field dependence of the room temperature tunneling magnetoresistance of Fe(30)/MgO(3)/Mn_3Sn(42) at the bias voltage of + 0.6 V (a) and − 0.5 V (b). The electric field and magnetic field are perpendicular (i.e. $E \perp B$) corresponding to Fig. 3c in the main text. The maximum magnetic field is 15 mT so that only Fe moments switch. Before scanning the small magnetic field, a large field of 1.5 T is applied to align the moment direction of Mn_3Sn. **c**, Bias dependence of the tunneling resistance of Fe(30)/MgO(3)/Mn_3Sn(42) from − 0.8 V to + 0.8 V. **d**, Bias dependence of the tunneling magnetoresistance ratio of Fe(30)/MgO(3)/Mn_3Sn(42) from − 0.8 V to + 0.8 V.

Extended Data Fig. 7 Bias dependent tunneling magnetoresistance of Fe/MgO/Mn_3Sn tunnel junctions having the perpendicular or parallel configuration between the electric field and magnetic field.

a, Schematic viewgraph indicating the relative position in energy for Mn_3Sn and Fe. The bar indicates the simplified energy dependence of the density of states of the band in Mn_3Sn and Fe. Tunneling conductance (red arrow) takes its maximum at the Fermi energy (E_F). **b**, Bias dependence of the tunneling resistance of the MTJs having the parallel ($E//B$,V(30)/Fe(0.6)/MgO(3)/Mn_3Sn(42)) and perpendicular ($E \perp B$, Fe(30)/MgO(3)/Mn_3Sn(42)) configurations, measured between − 0.8 V to 0.8 V. The horizontal broken line is a guide to the eye.

Extended Data Fig. 8 Tunneling magnetoresistance of Mn_3Sn/MgO/Mn_3Sn stacks with different MgO thickness.

a, Field dependence of the tunneling magnetoresistance of $Mn_3Sn(12)/MgO(3.3)/Mn_3Sn(42)$ at room temperature. The magnetic field and electric current are both applied perpendicular to the Mn_3Sn plane. **b**, Comparison between experimental tunneling magnetoresistance of $Mn_3Sn/MgO/Mn_3Sn$ and numerically calculated tunneling magnetoresistance of Mn_3Sn/vacuum/Mn_3Sn as a function of resistance area.

Extended Data Fig. 9 Energy dispersive X-ray spectroscopy mapping of $Mn_3Sn/MgO/Mn_3Sn$ stacks.

a-g, Elemental mappings for the area of Bright-STEM images (a) are shown in (b) Ta (yellow), (c) Sn (purple), (d) Mn (blue), (e) O (red), (f) Mg (green), and (g) W (orange). The white scale bars in (a-g) represent a length of 20 nm. Smooth W/Mn_3Sn and Mn_3Sn/MgO interfaces are clearly observed.

Extended Data Fig. 10 Transmission electron microscopy (TEM) images for the $Mn_3Sn/MgO/Mn_3Sn$ stack.

a. Cross-sectional TEM image of the heterostructure of the AFMTJ, MgO(001) substrate/W/Mn_3Sn/MgO/Mn_3Sn/Ta. **b—e**. Nano-beam electron diffraction patterns of the Mn_3Sn (top) (b), MgO (c), Mn_3Sn (bottom) (d), and W (e) layers. Measurement spots with ϕ 1 nm diameter are indicated as filled circles in the panel (a). Indexed spots are indicated by open circles and corresponding lattice planes are described next to the images. **f**. Representative crystal structure of Mn_3Sn in the Mn_3Sn bottom layer derived from the diffraction in the panel (d). Transparent light green plane and red plane indicate the plane parallel to the kagome (0001) plane and \((01\bar{1}1)\) plane, respectively. The unit cell of Mn_3Sn is denoted by a solid blue line. The viewpoint is the same as in the panels (a) and (d). **g**. Schematics of the atomic arrangement in the \((01\bar{1}0)\) plane of Mn_3Sn and (001) plane of W and MgO.

Supplementary information

Supplementary Information

Supplementary Figs. 1–5 and references.

Source data

Source Data Fig. 1

Source Data Fig. 2

Source Data Fig. 3

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Chen, X., Higo, T., Tanaka, K. *et al.* Octupole-driven magnetoresistance in an antiferromagnetic tunnel junction. *Nature* **613**, 490–495 (2023).

https://doi.org/10.1038/s41586-022-05463-w

- Received: 11 January 2022
- Accepted: 19 October 2022
- Published: 18 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05463-w

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05463-w

- Article
- Open Access
- Published: 18 January 2023

Vertical organic electrochemical transistors for complementary circuits

- Wei Huang ,
- Jianhua Chen,
- Yao Yao,
- Ding Zheng ,
- Xudong Ji,
- Liang-Wen Feng,
- David Moore,
- Nicholas R. Glavin,
- Miao Xie,
- Yao Chen,
- Robert M. Pankow,
- Abhijith Surendran,
- Zhi Wang,
- Yu Xia,
- Libing Bai,
- Jonathan Rivnay,
- Jianfeng Ping,
- Xugang Guo,
- Yuhua Cheng ,
- Tobin J. Marks &
- …
- Antonio Facchetti

Show authors

Nature **volume 613**, pages 496–502 (2023)

- 4203 Accesses
- 96 Altmetric
- Metrics details

Abstract

Organic electrochemical transistors (OECTs) and OECT-based circuitry offer great potential in bioelectronics, wearable electronics and artificial neuromorphic electronics because of their exceptionally low driving voltages (<1 V), low power consumption (<1 µW), high transconductances (>10 mS) and biocompatibility[1,2,3,4,5]. However, the successful realization of critical complementary logic OECTs is currently limited by temporal and/or operational instability, slow redox processes and/or switching, incompatibility with high-density monolithic integration and inferior n-type OECT performance[6,7,8]. Here we demonstrate p- and n-type vertical OECTs with balanced and ultra-high performance by blending redox-active semiconducting polymers with a redox-inactive photocurable and/or photopatternable polymer to form an ion-permeable semiconducting channel, implemented in a simple, scalable vertical architecture that has a dense, impermeable top contact. Footprint current densities exceeding 1 kA cm^{-2} at less than ±0.7 V, transconductances of 0.2–0.4 S, short transient times of less than 1 ms and ultra-stable switching (>50,000 cycles) are achieved in, to our knowledge, the first vertically stacked complementary vertical OECT logic circuits. This architecture opens many possibilities for fundamental studies of organic semiconductor redox chemistry and physics in nanoscopically confined spaces, without macroscopic electrolyte contact, as well as wearable and implantable device applications.

Main

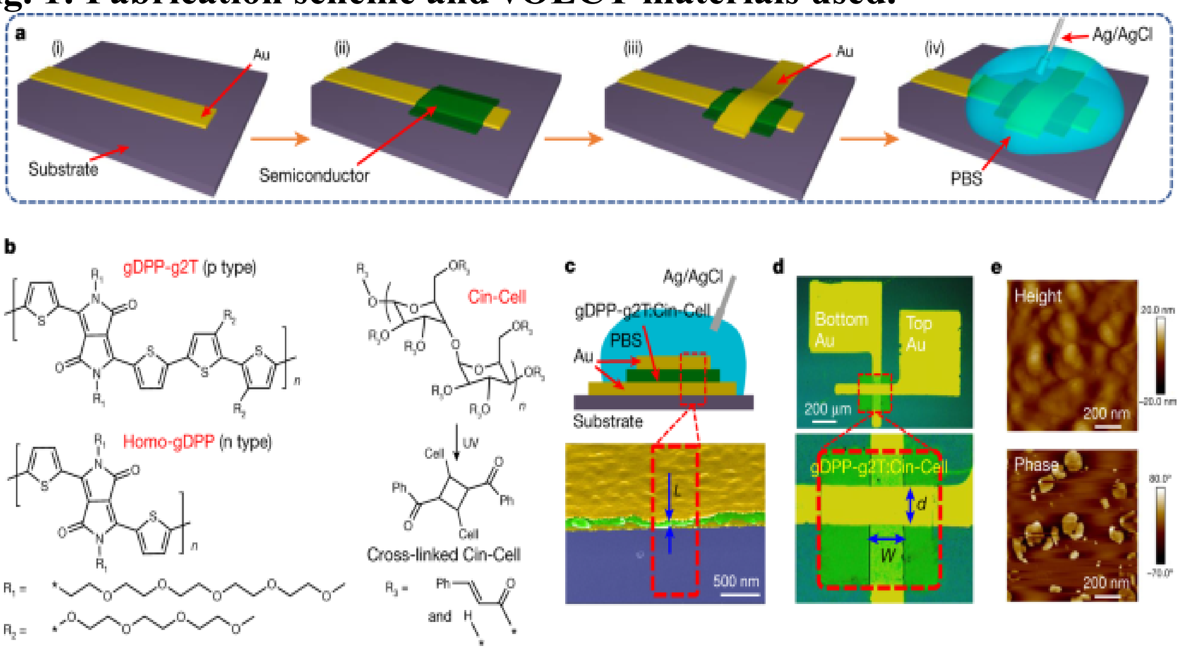
Organic electrochemical transistors (OECTs) are attractive for bioelectronics, wearable electronics and neuromorphic electronics because of their low driving voltage, low power consumption, high transconductance

and facile integration in mechanically flexible platforms[1,2,3,5,9,10,11]. However, further OECT advances face challenges. (1) Despite progress[8], poor electron-transporting (n-type) OECT performance versus their hole-transporting (p-type) counterparts (approximately 1,000 times lower transconductance and/or current density)[6,7,12], hinders the development of complementary logic and sensitivity to in vivo relevant analyte cations (for example, Na^+, K^+, Ca^{2+}, Fe^{3+} and Zn^{2+}) for biosensor development. (2) Temporal and/or operational instability hinders all possible applications. (3) Unbalanced p-type and n-type OECT performance prevents integration into complementary circuits[13,14]. (4) Slow redox processes lead to sluggish switching. (5) State-of-the-art conventional OECTs (cOECTs), having planar source–drain electrode architectures, require small channel lengths (L) of at most 10 µm, along with precisely patterned semiconducting layers and electrode coatings with passive materials, for high transconductance (g_m) and fast switching (approximately in the millisecond range)[15], requiring complex fabrication methodologies[15,16]. Note that conventional photolithography can only reliably realize features or L larger than 1 µm (ref. [16]), and although printing and laser cutting offer simplified cOECT fabrication, this is at the expense of performance[17,18,19]. Moreover, to increase g_m, OECTs typically use thick semiconducting films, inevitably compromising switching speeds because high g_m values require efficient ion exchange between the electrolyte and the bulk semiconductor[20]. Consequently, without progress in materials design, particularly for n-type semiconductors, and the realization of new device architectures, OECT applications will remain limited in scope.

In this report, we demonstrate high-performance p- and n-type OECTs and complementary circuits by using a vertical device architecture (vertical OECT, hereafter named vOECT) readily fabricated by thermal evaporation and masking of impermeable and dense Au source–drain electrodes and spin-coating and photopatterning of an ion-conducting semiconductor channel. The vOECT fabrication process is illustrated in Fig. 1a and details can be found in the Methods. The key to this process is the use of a redox-active p-type (gDPP-g2T) or n-type (Homo-gDPP) semiconducting polymer blended with a redox-inert and photocurable polymer component (cinnamate-cellulose polymer (Cin-Cell)) as the OECT channel (see the

structures in Fig. 1b, the synthesis process in the Methods and Extended Data Fig. 1). On the basis of the control experiments (vide infra) the optimal semiconducting polymer:Cin-Cell weight ratio was found to be 9:2. A vOECT geometry cross-section and selected optical and scanning electron microscopy (SEM) images (Fig. 1c,d) indicate that the channel length (*L*) is the semiconductor layer thickness (approximately 100 nm), the widths of the bottom and the top electrodes define the channel width (*W*) and the nominal depth (*d*) of the semiconductor, respectively. cOECTs and vOECTs that use polymers without ion-conducting ethylene glycol side chains were also fabricated as controls; their performance is marginal (Extended Data Fig. 2).

Fig. 1: Fabrication scheme and vOECT materials used.



a, Fabrication process for vOECTs: thermal evaporation of the bottom source electrode with a shadow mask (i), spin-coating and photopatterning of the semiconducting polymer + Cin-Cell blend (ii), thermal evaporation of the top drain electrode with a shadow mask (iii) and application of phosphate buffer solution (PBS) electrolyte and Ag/AgCl gate electrode (iv). **b**, Chemical structures of the redox-active semiconducting polymers (gDPP-g2T (p-type); Homo-gDPP (n-type)), and the redox-inactive cross-linkable polymer (Cin-Cell, cross-linking occurs through a photo-induced 2 + 2 cycloaddition reaction). **c**, Cross-section illustration of p-type vOECTs, along with a false-coloured cross-section SEM image showing the phase-separated layer sandwiched between two dense Au electrodes. **d**, Optical

image of a p-type vOECT, in which the electrode overlapping area is enlarged ($W = L = 70$ μm). **e**, AFM height and phase images of gDPP-g2T:Cin-Cell blends. In all samples, the gDPP-g2T:Cin-Cell weight ratio is 9:2.

Before evaluation of the device, the morphology and microstructure of the semiconducting polymer:Cin-Cell blend were characterized. As shown in Extended Data Fig. 3a,b, the pristine gDPP-g2T and HOMO-gDPP films are continuous and smooth (root-mean-square roughness, $\sigma_{r.m.s.} \approx 1$ nm), whereas both polymer blends with Cin-Cell are rougher after ultraviolet (UV) cross-linking/patterning ($\sigma_{r.m.s.} \approx 3$ nm) with evidence of phase separation in atomic force microscopy (AFM) (Fig. 1e), in which the Cin-Cell forms pillar-like structures that should enhance the structural robustness and stability. Thus, the Cin-Cell in the semiconducting matrix acts not only as a photopatterning component of the channel, but, most importantly, as an OECT structural stabilizer (vide infra). The two-dimensional grazing incidence wide-angle X-ray scattering (2D–GIWAXS, Extended Data Fig. 3c–e) patterns of the pure polymer and polymer:Cin-Cell mixture films are similar, corroborating phase separation and demonstrating that Cin-Cell addition does not substantially alter the overall film texturing and the polymer chain order.

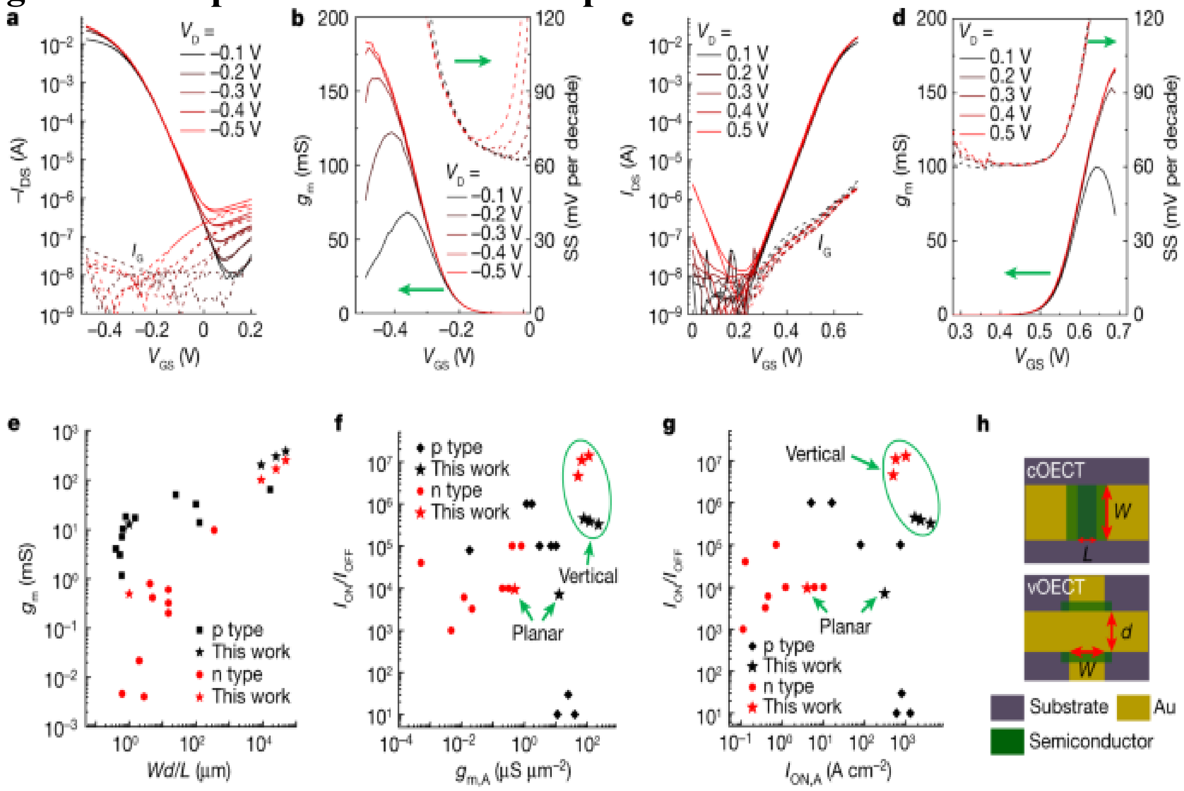
Next, the vOECTs and cOECTs were tested and the performance parameters were extracted following standard procedures (Extended Data Table 1)[7,21]. Before discussing the results, note that control vOECTs based on semiconducting polymers without ethylene glycol side chains exhibit a negligible transistor response, demonstrating that the hydrophilic ion-complexing polymer facilitates ion penetration across the nanoscopic thin electrolyte-blend interface (Extended Data Fig. 2b,c). Furthermore, the performance of control vOECTs with varying Cin-Cell weight contents indicate (Extended Data Fig. 2d–g) that the device yield without Cin-Cell is low and, most importantly, that such devices are unstable after very few repeated gate voltage (V_G) cycles, mainly reflecting top electrode delamination, whereas those that use a semiconducting polymer:Cin-Cell weight ratio of more than 9:2 exhibit poor performance (low current on (I_{ON}) and large hysteresis), because of reduced redox-active polymer content and constricted ion diffusion. Consequently, all of the data reported here are

for cross-linked semiconducting polymer:Cin-Cell blends with a 9:2 weight ratio.

The vOECT and cOECT transfer characteristics and the corresponding g_m–subthreshold swing (SS) plots (Fig. 2a–d and Extended Data Fig. 4) demonstrate extraordinary performances for both p- and n-type vOECTs, achieving maximum drain currents (I_{ON}) of $(8.2 \pm 0.5) \times 10^{-2}$ A (drain voltage (V_D) = −0.5 V, V_G = −0.5 V) and $(2.5 \pm 0.1) \times 10^{-2}$ A (V_D = +0.5 V, V_G = +0.7 V), and g_m values as high as 384.1 ± 17.8 mS and 251.2 ± 7.6 mS, respectively (Fig. 2e). Note that despite the ultra-small channel lengths ($L \approx$ 100 nm), the I_{ON}/current off (I_{OFF}) ratios of both devices are impressive ($\geq 10^6$), exclusively due to the higher I_{ON} and low I_{OFF}. All of the p- and n-type vOECTs retain stable turn-on voltages (V_{ON}) of +0.10 and +0.21 V as well as SS values of approximately 60 and approximately 62 mV per decade, respectively, upon scanning V_D from ±0.1 to ±0.5 V. More relevant parameters for vertical architectures are the area-normalized g_m ($g_{m,A}$) and I_{ON} ($I_{ON,A}$) metrics[22]. As shown in Fig. 2f,g, $g_{m,A}$ ($I_{ON,A}$) values as high as 226.1 μS μm^{-2} (4,036 A cm^{-2}) and 112.4 μS μm^{-2} (1,015 A cm^{-2}) are achieved for p- and n-type vOECTs, respectively. These values are about 18 times (13 times) and 100 times (1,000 times) greater than those measured in the corresponding p- and n-type cOECTs, respectively (Extended Data Table 1). Thus, the present p-type vOECTs exhibit, to our knowledge, the highest $g_{m,A}$ and $I_{ON,A}$ values reported up until now, even surpassing those of heavily doped and/or depletion-mode poly(3,4-ethylenedioxythiophene):polystyrene sulfonate (PEDOT:PSS) cOECTs. In addition, to our knowledge, the present n-type vOECT performance surpasses all previously reported OECTs (including p-type OECTs) in terms of $g_{m,A}$ and I_{ON}/I_{OFF}(refs. [7,8,14,15,21,23,24,25,26,27,28,29,30,31,32]). Importantly, the present vOECT structures also have reduced footprints because the contacting lines also function as source and drain contacts, eliminating the need for extra source–drain pads overlapping with the channel materials that are required in cOECTs. This vertical architecture and blending strategy with the Cin-Cell is also applicable to other mixed ionic-electronic semiconductors, with the vOECTs based on two p-type (Pg2T-T and PIBET-AO) and two n-type (polyethylene glycol-N2200 (PEG-N2200) and BTI2)

polymers exhibiting similarly enhanced transistor performance versus their planar counterparts (Extended Data Fig. 5).

Fig. 2: OECT performance and comparison with literature data.



a–**d**, Representative transfer characteristics (**a**,**c**) and corresponding g_m and SS curves (**b**,**d**) of p-type gDPP-g2T ($W = d = 30$ μm) (**a**,**b**) and n-type Homo-gDPP ($W = d = 50$ μm) (**c**,**d**) vOECTs. ($L \approx 100$ nm). **e**, g_m as a function of Wd/L for the present vOECTs and cOECTs as well as the previously reported OECTs[7,8,14,15,21,23,24,25,26,27,28,29,30,31,32]. Comparisons of current on/off ratio (I_{ON}/I_{OFF}) versus g_m per unit area ($g_{m,A}$) (**f**) and on-current per unit area ($I_{ON,A}$) (**g**) for different v- and cOECTs. Note, different asterisks are data of this work based on different W and d (Extended Data Table 1). **h**, Cartoon illustrating how the $g_{m,A}$ and $I_{ON,A}$ are calculated, where $g_{m,A} = g_m/(WL)$, $I_{ON,A} = I_{ON}/(WL)$ for cOECT, whereas $g_{m,A} = g_m/(Wd)$, $I_{ON,A} = I_{ON}/(Wd)$ for vOECT.

Source data

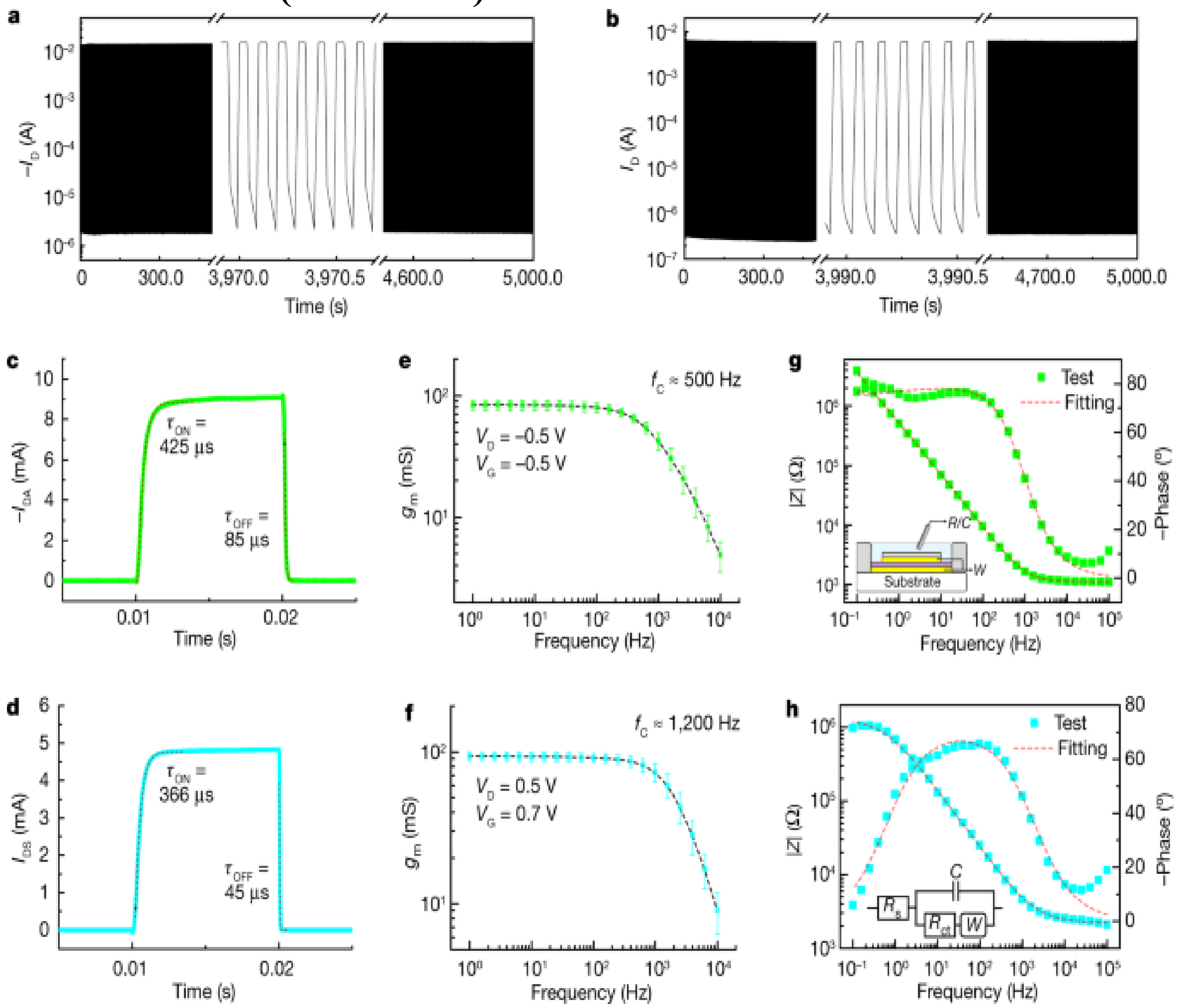
To our knowledge, there are no examples of truly vertical electrochemical devices in which transistor behaviour is observed throughout the entire semiconductor bulk by using ion-impermeable contacts. Previously reported pioneering organic transistor architectures with vertical source–drain arrangements functioned as OECTs only when using permeable (for example, Ag nanowires) electrodes[22,26,33], or when operated as electrical double-layer transistors or field-effect transistors. Therefore, bulk ion penetration and redox reactions are not involved, and only a small semiconductor volume under the top contact functions as the charge-carrying channel[34,35,36,37,38]. In contrast, the present approach uses simple thermal evaporation of dense and thick (150 nm) Au electrodes through a shadow mask, in combination with a photopatternable semiconductor layer, to create a structure with excellent ion intercalation. By using a semi-transparent Au top electrode (Extended Data Fig. 6a–c), Supplementary Video 1 demonstrates that the electrochromic switch associated with the redox chemistry encompasses the entire semiconductor area between the top and bottom electrodes and is not confined to the two narrow regions of the semiconductor that are in direct contact with the electrolyte. Evidence for bulk doping is further supported by the following observations: (1) I_{ON} and g_m of the vOECTs with different top electrode widths but identical bottom electrode widths, thus meaning an equal electrolyte–semiconductor interfacial area but a different semiconductor area or mass available for doping or de-doping, increase linearly (Extended Data Fig. 6d–f), illustrating that doping is not limited to the semiconductor–electrolyte interface. (2) Depletion-mode PEDOT:PSS vOECTs exhibit excellent switching behaviour and can be efficiently turned off (Extended Data Fig. 6g), which would be impossible if only interfacial redox chemistry was occurring in the vOECTs. (3) The measured saturation I_{ON} values of the present vOECTs would carry unreasonably large electrical current densities (>10^7 A cm^{-2}) if the channel were only few nanometres thick, as in typical electrical double-layer transistors. (4) Finally, devices based on very hydrophobic blends, which do not support ion intercalation across the nanoscopic interface, are non-functional (vide supra, Extended Data Fig. 2c).

The present vOECTs also exhibit good transistor behaviour even when operated at a V_D of only ±0.001 V (Extended Data Fig. 7a,b). Note,

especially for the n-type vOECTs, V_{ON} shifts from +0.43 to +0.21 V when V_D is only increased from +0.001 to +0.1 V because of the drain-induced barrier lowering, which is a short channel effect[22]. For n-type cOECTs reported in the literature, and here specifically for Homo-gDPP cOECT control (Extended Data Fig. 4h), the energetic mismatch between the n-type semiconductor LUMO level and the Au electrode work function results in a very high V_{ON} (>+0.4 V), and the limited electrochemical window of the aqueous electrolyte prevents the application of large V_G biases. This is one of the key limitations of current n-type cOECTs[39] and it is where drain-induced barrier lowering plays a key role in the n-type vOECT performance enhancement seen here. Common issues of short channel transistors, such as loss of saturation[40], V_T roll-off and reduced current modulation[22], which are equally as important, are absent in the vOECTs (Fig. 2 and Extended Data Fig. 7c,d). This result is possible only if the redox processes modulate the carrier concentration of the entire semiconducting layer[2,41]. The low SS of approximately 60 mV per decade measured for both vOECTs (Fig. 2b,d) provides more convincing proof of the extremely effective gating in the present vertical architecture. Furthermore, unlike cOECTs in which the region with SS approximately 60 mV per decade, if achieved, is narrow (Extended Data Fig. 4g,h), the present vOECTs have a very wide subthreshold region (0.0 ≈ −0.2 V for gDPP-g2T and +0.3 ≈ +0.6 V for Homo-gDPP) with SS near or equalling the approximately 60 mV per decade thermal limit. The wide subthreshold region is particularly useful for applications in which high voltage gain and low power consumption are vital[42,43].

The cycling stability along with the transient response of the OECTs were next assessed. As shown in Fig. 3a,b, for both p- and n-type vOECTs, more than 50,000 stable switching cycles are recorded, which is an order of magnitude higher than for literature values of OECTs, especially for n-type devices[21,44]. Note that the stability of the ubiquitous PEDOT:PSS in depletion-mode vOECTs is also greatly stabilized compared to that in cOECT architectures (Extended Data Fig. 7e,f). Furthermore, the vOECT turn-on transient time (τ_{ON}) is less than 0.5 ms for both devices (Fig. 3c,d), and is comparable to those of the corresponding precisely patterned cOECTs (Extended Data Fig. 7g,h).

Fig. 3: vOECT stability, switching time and frequency-dependent transconductance (bandwidth) characteristics.



a–**d**, Cycling stability (cycling frequency of 10 Hz) (**a**,**b**) and transient response (**c**,**d**) of p-type gDPP-g2T (**a**,**c**) and n-type Homo-gDPP (**b**,**d**) vOECTs, where $V_D = -0.1$ V, V_G is switching between 0 V and −0.5 V for the p-type vOECT, and $V_D = +0.1$ V, V_G is switching between 0 V and +0.7 V for the n-type vOECT. Note, for both vOECTs, $W = d = 30$ µm, $L \approx 100$ nm. **e**,**f**, Frequency-dependence of the small signal transconductance of p-type gDPP-g2T (**e**) and n-type Homo-gDPP (**f**) vOECTs, where the bias status is indicated in the figure and an additional 10 mV peak-to-peak gate voltage oscillation is applied (error bars represent s.d. for $n = 6$). **g**,**h**, EIS of vertical configuration (|Z|) ($W = d = 30$ µm, $L \approx 100$ nm) based on p-type gDPP-g2T:Cin-Cell (**g**) and Homo-gDPP:Cin-Cell (**h**). The insets in **g** and **h**

are the EIS measurement setup (**g**) and the equivalent circuit (2R1C behavior) (**h**).

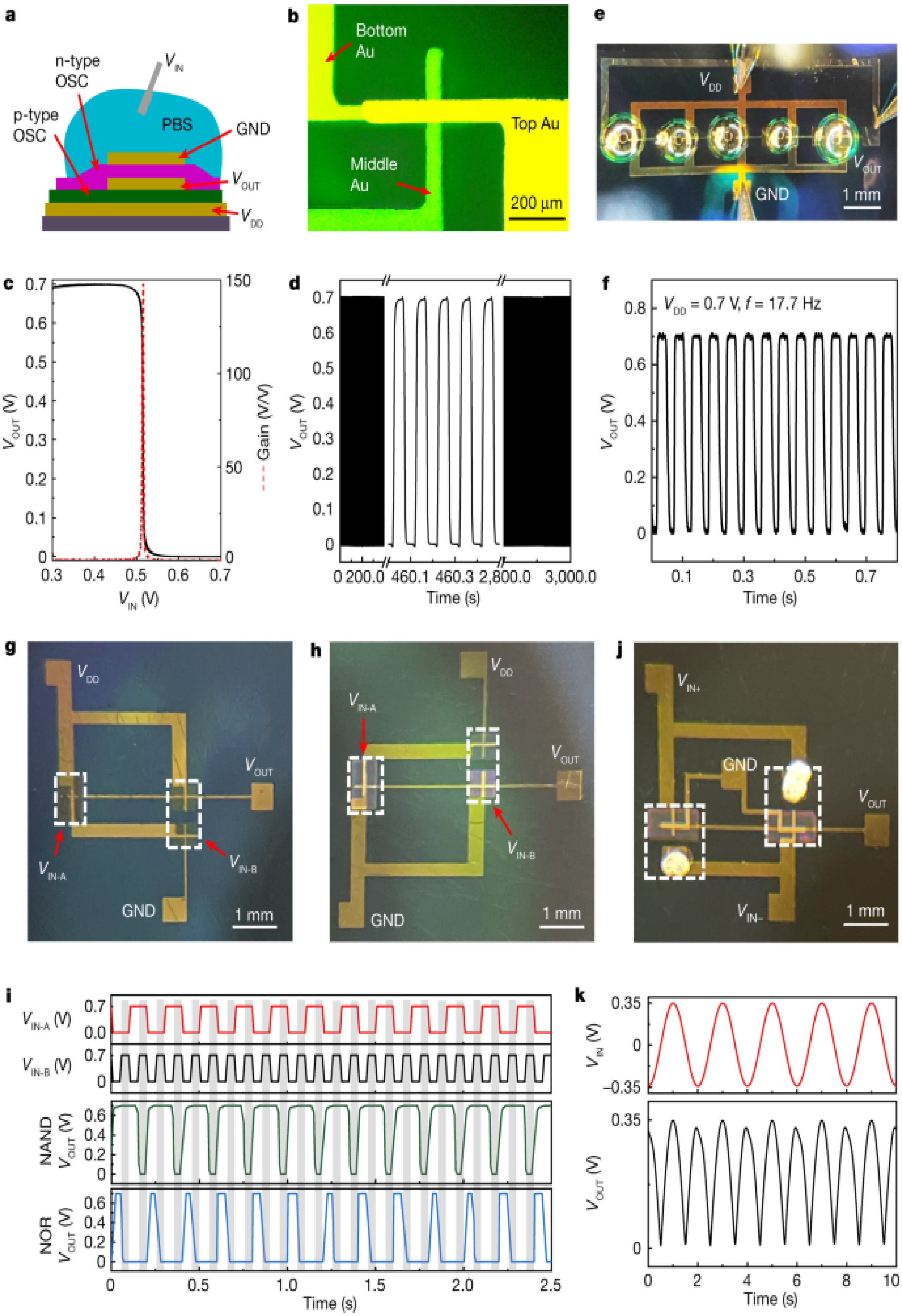
Source data

To validate the fast switching process and to understand the underlying mechanism, bandwidth and electrochemical impedance spectroscopy (EIS) measurements for both cOECTs and vOECTs were carried out (Fig. 3e,f and Extended Data Figs. 7i and 8). Cut-off frequencies (f_c) of approximately 500 and approximately 1,200 Hz are measured for p- and n-type vOECTs, respectively, which are consistent with the transient responses in Fig. 3c, whereas the EIS based on the vertical configuration also exhibits typical 2R1C behaviour as that in standard two-electrode configurations (Fig. 3g,h and Extended Data Fig. 8a–c). Moreover, vOECTs with different semiconductor film thicknesses were also fabricated and the transient responses were accessed (Extended Data Fig. 8d). As the film thickness (that is, the vOECT channel length) increases from 100 to 400 nm, τ_{ON} and τ_{OFF} increase from 425 µs and 85 µs (100 nm) to 32.6 ms and 966 µs (400 nm), respectively, showing that the confined electric field in the channel is the key for the fast transient response (Supplementary Video 2). On the basis of these results, it is fast bulk redox instead of interfacial doping near the semiconductor–electrolyte interface that is established, in which the short channel length leads to a strong electric field in the channel that effectively enhances the ion drift velocity, thereby leading to fast doping. Consequently, even in vOECTs for which the ion diffusion length is greater than 15 µm, the vOECT response times are amongst the shortest of the known n-type OECTs and are comparable to current state-of-the-art p-type OECTs, without extensive vOECT electrolyte or electrode patterning optimization[7,8,14,15,21,23,24,25,26,27,28,29,45].

So far, complementary logic has not been demonstrated for any type of vertical organic transistor architecture, mainly because of immaturity of the fabrication processes[22]. Here vertically stacked complementary inverters (VSCIs) are possible because of the unique vOECT operation mechanism, simple fabrication process and high stability of the present devices. Figure 4a shows a schematic of the VSCI, in which the n-type vOECT is located directly on top of the p-type vOECT. Such three-dimensional geometries

enable much higher integration densities that require a 50% smaller footprint per inverter (Fig. 4b). The voltage output characteristics indicate that the VSCI possesses a sharp voltage transition with a gain of up to approximately 150 (driving voltage = +0.7 V, Fig. 4c) and is stable for more than 30,000 switching cycles (Fig. 4d), further corroborating the excellent stability of both the n-type and p-type vOECTs. Thus, the present VSCI can also be used as an effective ion sensor over a wide concentration range (1 μM to approximately 0.1 M of KCl aqueous solution, Extended Data Fig. 8e,f), for which the transition voltage of the inverter can be effectively modulated to near half of voltage drain drain, $V_{DD}/2$.

Fig. 4: Vertically stacked complementary circuits based on vOECTs.


a
n-type OSC
V_IN
PBS
p-type OSC
GND
V_OUT
V_DD
b
Bottom Au
Top Au
Middle Au
200 μm
e
V_DD
V_OUT
GND
1 mm
c
V_OUT (V)
Gain (V/V)
V_IN (V)
0.3
0.4
0.5
0.6
0.7
0
0.1
0.2
150
100
50
d
Time (s)
0 200.0
460.1
460.3
2,800.0
3,000.0
f
V_DD = 0.7 V, f = 17.7 Hz
Time (s)
0.1
0.3
0.5
0.7
0.8
g
V_DD
V_OUT
V_IN-A
V_IN-B
GND
1 mm
h
V_DD
V_IN-A
V_OUT
V_IN-B
GND
1 mm
j
V_IN+
GND
V_OUT
V_IN−
1 mm
i
V_IN-A (V)
V_IN-B (V)
NAND V_OUT (V)
NOR V_OUT (V)
0.0
1.0
1.5
2.0
2.5
Time (s)
k
V_IN (V)
0.35
−0.35
V_OUT (V)
0
2
4
6
8
10
Time (s)

a, Illustration of a VSCI based on vOECTs (OSC = organic semiconductor). **b**, Top view of the VSCI, for which the Au electrode locations are indicated. **c**, Voltage output characteristics of the VSCI, along with the voltage gain. **d**, Switching stability of the VSCI with a switching frequency of 10 Hz. **e**,**f**, Photograph of a five-stage ring oscillator (**e**) and the corresponding output characteristics (**f**). **g**–**i**, Photograph of NAND (**g**) and NOR (**h**) circuits, and the corresponding voltage input/output characteristics (**i**). **j**, Photograph of a rectifier and the corresponding output characteristics (**k**). Note, in photographs **g**,**h** and **i**, the electrolyte and Ag/AgCl are omitted to provide a better view of the channel areas.

Source data

In addition, a five-stage ring oscillator was fabricated based on the present VSCI (Fig. 4e and Extended Data Fig. 9), and the output signal begins to oscillate between 0.0 and +0.7 V at a frequency of 17.7 Hz (V_{DD} = +0.7 V, Fig. 4f). This corresponds to a propagation delay of approximately 5.6 ms for each inverter. Finally, NAND and NOR logic gates operating between 0.0 and +0.7 V (Fig. 4g–i and Extended Data Fig. 9) as well as a VSCI-based rectifier (0.35 V amplitude, Fig. 4j,k and Extended Data Fig. 9) were fabricated, demonstrating a versatile library of circuitry elements. Note that previous cOECT-based ring oscillators, NANDs and NORs were fabricated with unipolar p-type cOECTs[17,46,47,48], whereas complementary circuits were limited at the preliminary stage of an inverter because of the low performance of the n-type cOECTs[12,14]. Thus, the present vOECTs enable not only VSCIs, which far outperform the corresponding state-of-the-art cOECT inverters[17,49,50], but they also facilitate the integration of this electrochemical technology into more complex complementary electronics.

In summary, this work reports vOECTs that demonstrate unprecedented performances for both p- and n-type operation modes. The device architecture demonstrated here is enabled by the synthesis of new electro-active and ion-permeable semiconducting polymers and by the interface engineering of electro-active blend layers. The devices are accessible by conventional fabrication processes and provide high fidelity and stable performance characteristics. They open up opportunities for fundamentally new system designs in diverse applications including low-cost diagnostics,

brain-machine interfaces, implantable and wearable devices, prosthetics and intelligent soft robotics, for which small effective footprints along with high g_m and low driving voltage metrics are essential requirements. Moreover, vOECTs offer a new design paradigm for flexible and stretchable complementary devices and related logic circuits.

Methods

Materials synthesis

The synthetic route to the new polymers gDPP-g2T and Homo-gDPP is illustrated in Extended Data Fig. 1a. Unless otherwise stated, all reactions were carried out under argon and the solvents were used without any purification. The reagents 2,5,8,11,14-pentaoxahexadecan-16-yl 4-methylbenzenesulfonate (**1**)[51], 3,6-di(thiophen-2-yl)-2,5-dihydropyrrolo[3,4-c]pyrrole-1,4-dione (**2**)[52] and 5,5′-bis(trimethyltin)-3,3′-bis(2-(2-(2-methoxyethoxy)ethoxy)ethoxy)-2,2′-bithiophene (**5**)[53] were synthesized according to previously reported procedures. Hexabutyldistannane (**6**) was purchased from Sigma-Aldrich. The Cin-Cell ploymer was prepared according to our previous publication[54].

Synthesis of 2,5-di(2,5,8,11,14-pentaoxahexadecan-16-yl)-3,6-di(thiophen-2-yl)-2,5-dihydropyr-rolo[3,4-c]pyrrole-1,4-dione (3)

Compound **1** (6.00 g, 14.76 mmol), compound **2** (1.84 g, 6.15 mmol), K_2CO_4 (4.25 g, 30.75 mmol) and 40 ml of dimethylformamide were added to a 100 ml single-neck round-bottom flask. The reaction mixture was purged with argon for 15 min and was then heated to 150 °C overnight. After cooling to 25 °C, the solvent was removed under reduced pressure. The residue was next dissolved in chloroform and was then washed with water and brine 3 times each. The organic phase was then dried over anhydrous Na_2SO_4, filtered and the solvent was removed under vacuum to leave the crude product, which was then purified by silica gel chromatography, eluting with chloroform/methanol (100:1 to 20:1). Compound **3** was obtained as a red solid (1.71 g; yield, 36%). ^{1}H NMR (500 MHz, $CDCl_3$, Extended Data Fig.

1b): δ (ppm) = 8.75 (d, J = 3.9 Hz, 2H), 7.64 (d, J = 5.0 Hz, 2H), 7.26 (dd, J = 5.0 Hz, 3.9 Hz, 2H), 4.26 (t, J = 6.3 Hz, 4H), 3.78 (t, J = 6.4 Hz, 4H), 3.66–3.51 (m, 32H), 3.36 (s, 6H). ^{13}C NMR (126 MHz, $CDCl_3$, Extended Data Fig. 1c): δ (ppm) = 161.54, 140.44, 134.78, 130.91, 129.68, 128.46, 107.89, 71.94, 70.72, 70.62, 70.60, 70.58, 70.57, 70.52, 68.94, 59.04, 41.88. High-resolution mass spectrometry (HRMS) matrix-assisted laser desorption–ionization (MALDI): calcd for $C_{36}H_{52}N_2NaO_{12}S_2$ (M + Na^+): 791.2859; found, 791.2851.

Synthesis of 3,6-bis(5-bromothiophen-2-yl)-2,5-di(2,5,8,11,14-pentaoxahexadecan-16-yl)-2,5-dihydropyrrolo[3,4-c]pyrrole-1,4-dione (4)

Compound **3** (1.00 g, 1.30 mmol) was dissolved in 30 ml of chloroform in a 100 ml single-neck round-bottom flask. The reaction mixture was cooled to 0 °C and *N*-bromosuccinimide (0.48 g, 2.73 mmol) was added in one portion under argon. The reaction mixture was slowly warmed to room temperature and was stirred overnight in the dark. Water (100 ml) was added and the resutling solution was stirred for 30 min. The organic layer was separated and was dried over anhydrous Na_2SO_4, filtered and the solvent was removed under vacuum to leave a residue that was purified by silica gel chromatography with chloroform/methanol (100:1 to 50:1). Compound **4** was obtained as a purple solid (0.86 g, yield 71%). ^{1}H NMR (500 MHz, $CDCl_3$, Extended Data Fig. 1d): δ (ppm) = 8.48 (d, J = 4.2 Hz, 2H), 7.20 (d, J = 4.2 Hz, 2H), 4.16 (t, J = 6.0 Hz, 4H), 3.76 (t, J = 6.0 Hz, 4H), 3.66–3.51 (m, 32H), 3.36 (s, 6H). ^{13}C NMR (126 MHz, $CDCl_3$, Extended Data Fig. 1e): δ (ppm) = 161.26, 139.48, 134.86, 131.41, 131.12, 119.35, 107.97, 71.93, 70.76, 70.61, 70.58, 70.56, 70.50, 68.94, 59.03, 42.24, 29.60. HRMS (MALDI): calcd for $C_{36}H_{50}Br_2N_2NaO_{12}S_2$ (M + Na^+): 949.1049; found, 949.1044.

Synthesis of polymer gDPP-g2T

Compound **4** (92.67 mg, 0.10 mmol), compound **5** (81.62 mg, 0.1 mmol), $Pd_2(dba)_3$ (3.00 mg) and $P(o\text{-tol})_3$ (7.60 mg) were added to a 10 ml reaction

vessel. After the reaction mixture was pump–purged for three cycles with argon, anhydrous toluene (1.5 ml) and dimethylformamide (1.5 ml) were added. The sealed vessel was next heated at 110 °C for 12 h. The polymer was then end-capped with 20 µl of 2-(tributylstannyl)-thiophene and then 50 µl of 2-bromothiophene, with each step being carried out at 110 °C for 1 h. After cooling to room temperature, the mixture was poured into 100 ml of MeOH + 1 ml of concentrated HCl. The resulting precipitate was collected by filtration and then purified by Soxhlet extraction using methanol, acetone, hexane and then chloroform. The chloroform portion was concentrated and then poured into MeOH (approximately 100 ml). The resulting precipitate was collected by vacuum filtration as a black solid (106.51 mg, yield 86%). ^1H NMR (500 MHz, $C_2D_2Cl_4$, Extended Data Fig. 1f): δ (ppm) = 8.73–8.69 (br, 2H), 7.43–6.84 (br, 4H), 4.34–4.25 (br, 8H), 3.94–3.45 (m, 56H), 3.29 (s, 12H). Anal. calcd for $[C_{58}H_{84}N_2O_{20}S_4]_n$: C, 55.40; H, 6.73; N, 2.23. Found: C, 55.41; H, 6.67; N, 2.37.

Synthesis of polymer Homo-gDPP

The polymer Homo-gDPP was synthesized by using the same method as used for gDPP-g2T. Compound **4** (199.00 mg, 0.21 mmol), compound **6** (124.57 mg, 0.21 mmol), $Pd_2(dba)_3$ (5.00 mg) and $P(o\text{-tol})_3$ (13.00 mg) were used as starting materials. The pure polymer was obtained as a black solid (100.00 mg, yield 62%). ^1H NMR (500 MHz, $C_2D_2Cl_4$, Extended Data Fig. 1g): δ (ppm) = 8.91–8.69 (br, 2H), 7.43–7.22 (br, 2H), 4.25 (br, 4H), 3.78–3.54 (br, 36H), 3.46 (br, 6H). Anal. calcd for $[C_{36}H_{52}N_2O_{12}S_2]_n$: C, 56.23; H, 6.82; N, 3.64. Found: C, 56.26; H, 6.83; N, 3.76.

Materials characterization

The ^1H and ^{13}C NMR spectra of the intermediates were recorded on a Bruker Ascend 500 MHz spectrometer by using deuterochloroform ($CDCl_3$) as the solvent at room temperature. The ^1H spectra of the polymers were recorded on a Bruker Ascend 500 MHz spectrometer by using dideutero-1,1,2,2-tetrachloroethane ($C_2D_2Cl_4$) at 100 °C, which was also used to

estimate the molecular weight. The purity of the polymers was verified by elemental analysis carried out at Midwest Microlabs Inc.

VT–NMR and M_n estimation by end-group analysis

The solutions for the NMR experiments were prepared by dissolving approximately 5 mg of polymer in 0.7 ml of $C_2D_2Cl_4$. The solutions were heated at 100 °C for 16 h before the measurements were taken to ensure complete dissolution of the polymer. The measurements were performed on a 400 MHz Bruker Avance III HD Nanobay at 100 °C, and the spectra were referenced to C_2DHCl_4 at 5.90 ppm. End-groups were identified based on literature compounds of similar structure[55,56]. Calculation of the M_n from the end-group analysis is based on equation (1), which is described in the literature[57].

$${n}_{x}=\frac{{a}_{x}{m}_{y}{n}_{y}}{{a}_{y}{m}_{x}} \quad (1)$$

where a_x is the corrected number of repeat unit protons, m_y is the number of end-group protons used for the calculation, a_y is the area of the end-group protons and m_x is he number of repeat unit protons.

For Homo-gDPP: $n_x = [(10.65)(2)(2)]/[(1)(2)] = 21.3 \approx 21$ and $M_n = (21 \times 0.76893) = 16.4$ kDa.

For gDPP-g2T: $n_x = [(16.8)(2)(2)]/[(1)(2)] = 33.6 \approx 34$ and $M_n = (34 \times 1.16744) = 39.7$ kDa.

OECT and complementary circuit fabrication

Semiconductor solution preparation

The gDPP-g2T, Homo-gDPP and Cin-Cell were first dissolved in chloroform at a concentration of 20 mg ml^{-1} and were filtered through a 0.45 μm polyvinylidene difluoride filter. Then, the gDPP-g2T or Homo-

gDPP solution was mixed with the Cin-Cell solution in a volume ratio of 9:2 for device fabrication by using the blends. For Pg2T-T, PIBET-AO, PEG-N2200 and BTI2, they were also first dissolved in chloroform at a concentration of 20 mg ml^{-1} and were filtered through a 0.45 µm polyvinylidene difluoride filter, then mixed with the Cin-Cell solution in a volume ratio of 9:2. For PEDOT:PSS (Xi'an Polymer Light Technology Corp.), a solution containing 1 ml of PH1000 (solid content approximately 1.3%, PEDOT content approximately 0.37%), 1 polyethylene glycol dimethacrylate (1.2:1 weight ratio versus PEDOT), 5 wt% of Irgacure 2959 (versus polyethylene glycol dimethacrylate) and 1 wt% of Capstone FS-30 (versus PH1000)[58] was prepared.

Conventional OECT fabrication

A Si wafer with a 300-nm-thick SiO_2 layer was used as the substrate. It was ultrasonically cleaned, first in an isopropyl alcohol bath for 20 min and then with oxygen plasma for 5 min. The S1813 photoresist was spin-coated at 4,000 rpm for 45 s, followed by annealing at 110 °C for 60 s and was then exposed under a maskless aligner system (MLA150; Heidelberg Instruments), developed in AZ400k (Microchemicals) for 40 s, rinsed with deionized water and blow-dried. Next, 3 nm of Cr and 50 nm of Au were deposited by thermal evaporation and were developed by soaking in acetone for 5 min to remove the S1813. Here, the patterned planar Au source–drain electrodes defined the channel dimension of $W = 100$ µm and $L = 10$ µm. The p- or n-type semiconductor blend solution was then spin-coated at 3,000 rpm for 20 s and was UV cross-linked for 30 s (Inpro Technologies F300S). At last, a droplet (approximately 1–20 µl, based on the channel area) of phosphate buffer solution (PBS, 1×) was applied onto the electrode overlapping area, and an Ag/AgCl electrode was inserted in the droplet acting as the OECT gate electrode. For the well-patterned cOECTs, the fabrication process can be found in the literature[59], and the devices have patterned semiconductor areas (100×20 μm^2, in which the channel length is 10 µm, the channel width is 100 µm, the channel thickness is 100 nm and the overlap with the source and/or drain is 5 µm on each side) and encapsulated source–drain electrodes.

Vertical OECT fabrication

An illustration of the vOECT fabrication process can be also found in Fig. 1a. The vOECTs were also fabricated on a pre-cleaned Si/300 nm SiO_2 wafer. First, 3 nm of Cr and 150 nm of Au (rate approximately 0.5–2.0 Å s^{-1}) were thermally evaporated with a shadow mask as the bottom source electrode. Next, the semiconductor blend solution was spin-coated on the substrate at 3,000 rpm for 20 s. The semiconducting layer was then UV cross-linked for 30 s (Inpro Technologies F300S). Note that the semiconducting layer can be further patterned by developing it in chloroform for 3 s and blow-drying if cross-linked with a photomask. The top drain electrode (150 nm Au) was then thermally evaporated (rate approximately 0.5–2.0 Å s^{-1}) with a shadow mask while maintaining the substrate at a temperature of approximately 20 °C with a back water-cooling system. Finally, a droplet (approximately 1–20 µl, based on the channel area) of PBS (1×) was applied on the electrode overlapping area, and an Ag/AgCl electrode was inserted in the droplet acting as the OECT gate electrode. Control devices using the pure semiconductors were fabricated following the same procedure but by using pure polymer solutions and without UV exposure.

Complementary inverter fabrication

An illustration of the fabrication process can be found in Extended Data Fig. 9a. For the inverter fabrication, a layer of the opposite type of semiconductor blend was spin-coated (3,000 rpm for 20 s) directly onto the first vOECT (before applying the PBS electrolyte and the Ag/AgCl electrode), and was UV cross-linked for 30 s. Next, the third Au electrode (150 nm) was evaporated with a shadow mask as describe above. Note that the third Au electrode was carefully aligned to overlap with the active area of the bottom vOECT. Finally, a droplet (approximately 1–20 µl, based on the channel area) of PBS (×1) was applied on the electrode overlapping area, and an Ag/AgCl electrode was inserted in the droplet acting as V_{IN} of the inverter.

Complementary ring oscillator fabrication

An illustration of the fabrication process can be found in Extended Data Fig. 9b. The five-stage ring oscillator was also fabricated on a pre-cleaned Si/300 nm SiO_2 wafer. First, 3 nm Cr and 150 nm Au were thermally evaporated with a shadow mask as the bottom electrode (V_{DD}). Next, the p-type gDPP-g2T:Cin-Cell mixture solution was spin-coated on the substrate at 3,000 rpm for 20 s and was cross-linked under UV light for 30 s with a shadow mask. The film was patterned by immersing it in chloroform for 3 s and blow-drying. Next, 150 nm Au were thermally evaporated with a shadow mask as the middle electrode (V_{OUT}). The n-type Homo-gDPP:Cin-Cell mixture was then spin-coated and photopatterned in the same way as for the p-type polymer blend. Then, 150 nm Au top electrode (ground, GND) was thermally evaporated with a shadow mask. Pure Cin-Cell solution was then spin-coated at 5,000 rpm for 20 s, cross-linked under UV light for 60 s with a shadow mask and developed in chloroform for 3 s, to leave openings for the active channel areas and V_{OUT} electrodes. A Ag/AgCl paste (Creative Materials, 125-20) was applied on the V_{OUT} electrodes of each inverter and vacuum dried for 30 min. Finally, a drop of PBS electrolyte (approximately 2 µl) was applied on each V_{OUT} electrode and its adjacent inverter active channel area.

NAND and NOR fabrication

An illustration of the fabrication process can be found in Extended Data Fig. 9c,d. NAND and NOR logic gates were also fabricated on a pre-cleaned Si/300 nm SiO_2 wafer. First, 3 nm Cr and a 150 nm Au were thermally evaporated with a shadow mask as the bottom electrode. Next, the p-type gDPP-g2T:Cin-Cell mixture solution was spin-coated on the substrate at 3,000 rpm for 20 s and was cross-linked under UV light for 30 s with a shadow mask. The film was patterned by immersing in chloroform for 3 s and blow-dried. Next, 150 nm Au were thermally evaporated with a shadow mask as the middle electrode (V_{OUT}). The n-type Homo-gDPP:Cin-Cell mixture was then spin-coated and photopatterned in the same way as for the p-type polymer blend but with a different shadow mask. Then, 150 nm Au top electrode was thermally evaporated with a shadow mask. Pure Cin-Cell solution was spin-coated at 5,000 rpm for 20s, cross-linked under UV light for 60 s with a shadow mask and developed in chloroform for 3 s, to leave

openings for the active channel areas. Finally, two drops of PBS electrolyte (approximately 2 μl) were applied on each V_{IN} area along with two Ag/AgCl electrodes as $V_{IN\text{-}A}$ and $V_{IN\text{-}B}$, respectively.

Rectifier fabrication

An illustration of the fabrication process can be found in Extended Data Fig. 9e. The rectifier was also fabricated on a pre-cleaned Si/300 nm SiO_2 wafer. First, 3 nm Cr and 150 nm Au were thermally evaporated with a shadow mask as the bottom electrode (V_{OUT}). Next, the p-type gDPP-g2T:Cin-Cell mixture solution was spin-coated on the substrate at 3,000 rpm for 20 s and was cross-linked under UV light for 30 s with a shadow mask. The film was patterned by immersing it in chloroform for 3 s and blow-drying. Next, 150 nm Au were thermally evaporated with a shadow mask as middle electrode (V_{IN+} and $V_{IN–}$). The n-type Homo-gDPP:Cin-Cell mixture was then spin-coated and photopatterned as that of the p-type polymer blend. Then, 150 nm Au top electrode (GND) was thermally evaporated with a shadow mask. Pure Cin-Cell solution was then spin-coated at 5,000 rpm for 20 s, cross-linked under UV light for 60 s with a shadow mask and developed in chloroform for 3 s to leave openings for the active channel areas and V_{IN} electrodes. A Ag/AgCl paste (Creative Materials, 125-20) was applied on the V_{IN} electrodes and was vacuum dried for 30 min. Finally, two drops of PBS electrolyte (approximately 2 μl) were applied on each V_{IN} electrode and its adjacent active channel area.

Device characterization

Transistor measurement

The electrical characterization of the OECTs and inverters was carried with an Agilent B1500A semiconductor parameter analyser in ambient conditions. The voltage sweeping speed was 0.1 V s^{-1} for the OECT measurements. For the transistor and inverter cycling tests, the voltage pulse was generated by a Keysight waveform generator (33500B), whereas the current–voltage variation was monitored with an Agilent B1500A. During

the cycling tests, to maintain a relatively stable PBS electrolyte concentration, a PDMS mould was placed on top of the device active area to confine the electrolyte displacement and to slow water evaporation. Transient time measurements were carried out with an FS-Pro (PDA) semiconductor parameter analyser. For the ring oscillator characterization, a constant V_{DD} of +0.7 V was applied with an Agilent B1500A, and the V_{OUT} was monitored by an oscilloscope (Tektronix, TDS 2014). For NAND and NOR characterization, square pulses (from 0.0 to ±0.7 V) with a frequency of 5 Hz and 10 Hz were applied as $V_{IN\text{-}A}$ and $V_{IN\text{-}B}$, respectively, by a Keysight waveform generator (33500B), and V_{OUT} was monitored by an Agilent B1500A. For rectifier characterization, two sinusoidal V_{IN} (V_{IN+} and V_{IN-} have a phase difference of 180°) with an amplitude of 0.35 V were generated by a Keysight waveform generator (33500B), and the V_{OUT} was monitored by an Agilent B1500A. All measurements were carried out in ambient conditions.

EIS measurements

All measurements were conducted by using a PalmSens4 potentiostat (PalmSens) with an Ag/AgCl pellet (Warner Instruments) as the reference and counter electrode, and a gold electrode coated with active materials as the working electrode. For measurements on the vertical structure, details can be found in Extended Data Fig. 7, which were performed in PBS (1×) electrolyte with a direct current offset as 0.5 V (for the p-type material) and −0.7 V (for the n-type material), superimposed by a 10 mV alternating current (a.c.) oscillation. The frequency of the a.c. oscillation ranges from 0.1 to 10^5 Hz.

Bandwidth measurements

Bandwidth measurements were conducted by accessing the g_m of the OECT as a function of the frequency of the gate voltage oscillation. The National Instruments (NI) SMU unit (NI PXIe-4143) was used for sourcing and measuring the drain–source voltage and current, as well as the gate current. The gate voltage was applied by using the data acquisition (DAQ) card from

NI (NI PXIe-6363)] and was measured with a NI BNC-2110. During the measurement, V_{DS} is equal to −0.5 V (p-type) and 0.5 V(n-type), whereas V_G is equal to −0.5 V (p-type) and 0.7 V (n-type), superimposed by a 10 mV a.c. oscillation. The frequency of the a.c. oscillation ranges from 1 to 10^4 Hz. All measurements were automated by using a custom LabVIEW programme (NI) and the data were processed by using the MATLAB software (Mathworks).

Semiconductor film characterization

SEM characterizations were carried out on a Hitachi SU8030 FE-SEM. AFM characterizations were acquired with a Bruker ICON System. GIWAXS measurements were performed at Beamline 8-ID-E1 at the Advanced Photon Source (APS) at Argonne National Laboratory. Samples were irradiated with a 10.9 keV X-ray beam at an incidence angle 0.125° to 0.135° in a vacuum for two summed exposures of 2.5 s (totalling 5 s of exposure), and scattered X-rays were recorded by a Pilatus 1 M detector located 228.16 mm from the sample at two different heights.

Discussion of the calculated mobilities of cOECT and vOECTs

The carrier mobilities of both the gDPP-g2T:Cin-Cell and Homo-gDPP:Cin-Cell were measured in vertical and planar OECTs[59]. For the planar or conventional architecture, the carrier mobility of the p-type cOECT is found to be 1.69 ± 0.19 cm^2 V^{-1} s^{-1}, which is comparable to other high-performance p-type OECTs. The n-type cOECT exhibits a high mobility of 0.13 ± 0.03 cm^2 V^{-1} s^{-1}, which is among the highest reported up until now. However, for the vertical devices, the calculated carrier mobilities are much lower, $(3.33 \pm 0.27) \times 10^{-3}$ cm^2 V^{-1} s^{-1} and $(3.06 \pm 0.61) \times 10^{-3}$ cm^2 V^{-1} s^{-1} for the p-type and n-type vOECTs, respectively.

The much lower and similar carrier mobilities in the vOECTs probably originate from the large series resistance from the source–drain electrodes and, to a lesser extent, the non-optimal polymer morphology of the spin-coated films, which typically enhances in-plane rather than out-of-plane organic semiconductor charge transport. Because the measured channel resistance in the vertical structure in the on-state is less than 10 Ω, series

resistances originating at the electrode–semiconductor interface and within the electrode contact will reduce the measured drain current notably[16]. Overall, these observations indicate that further optimization of charge injection and connecting-line conductivity, and the use of organic semiconductors that favour vertical charge transport as in those for organic photovoltaics, will probably enhance the current densities even further. Nevertheless, to properly evaluate the true carrier mobilities for the unconventional vertical architecture reported here, additional modelling and simulation efforts are required. This would be of great interest to the entire community.

Data availability

Source data are provided with this paper. Additional data related to this work are available from the corresponding authors upon request.

References

1. Park, S. et al. Self-powered ultra-flexible electronics via nano-grating-patterned organic photovoltaics. *Nature* **561**, 516–521 (2018).

2. Rivnay, J. et al. Organic electrochemical transistors. *Nat. Rev. Mater.* **3**, 17086 (2018).

3. van de Burgt, Y. et al. A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nat. Mater.* **16**, 414–418 (2017).

4. Khau, B. V., Scholz, A. D. & Reichmanis, E. Advances and opportunities in development of deformable organic electrochemical transistors. *J. Mater. Chem. C* **8**, 15067 (2020).

5. Ji, X. et al. Mimicking associative learning using an ion-trapping non-volatile synaptic organic electrochemical transistor. *Nat. Commun.* **12**, 2480 (2021).

6. Zeglio, E. & Inganas, O. Active materials for organic electrochemical transistors. *Adv. Mater.* **30**, e1800941 (2018).

7. Ohayon, D. et al. Influence of side chains on the n-type organic electrochemical transistor performance. *ACS Appl. Mater. Interfaces* **13**, 4253–4266 (2021).

8. Giovannitti, A. et al. N-type organic electrochemical transistors with stability in water. *Nat. Commun.* **7**, 13066 (2016).

9. Kim, J. H., Kim, S. M., Kim, G. & Yoon, M. H. Designing polymeric mixed conductors and their application to electrochemical-transistor-based biosensors. *Macromol. Biosci.* **20**, e2000211 (2020).

10. Spyropoulos, G. D., Gelinas, J. N. & Khodagholy, D. Internal ion-gated organic electrochemical transistor: a building block for integrated bioelectronics. *Sci. Adv.* **5**, eaau7378 (2019).

11. Yang, A. et al. Fabric organic electrochemical transistors for biosensors. *Adv. Mater.* **30**, e1800051 (2018).

12. Sun, H. D., Gerasimov, J., Berggren, M. & Fabiano, S. n-Type organic electrochemical transistors: materials and challenges. *J. Mater. Chem. C* **6**, 11778 (2018).

13. Romele, P. et al. Multiscale real time and high sensitivity ion detection with complementary organic electrochemical transistors amplifier. *Nat. Commun.* **11**, 3743 (2020).

14. Sun, H. et al. Complementary logic circuits based on high-performance n-type organic electrochemical transistors. *Adv. Mater.* **30**, 1704916 (2018).

15. Khodagholy, D. et al. High transconductance organic electrochemical transistors. *Nat. Commun.* **4**, 2133 (2013).

16. Donahue, M. J. et al. High-performance vertical organic electrochemical transistors. *Adv. Mater.* **30**, 1705031 (2018).

17. Andersson Ersman, P. et al. All-printed large-scale integrated circuits based on organic electrochemical transistors. *Nat. Commun.* **10**, 5053 (2019).

18. Rashid, R. B., Ciechowski, R. J. & Rivnay, J. Self-aligned, laser-cut organic electrochemical transistors. *Flex. Print. Electron.* **5**, 014007 (2020).

19. Schmatz, B., Lang, A. W. & Reynolds, J. R. Fully printed organic electrochemical transistors from green solvents. *Adv. Funct. Mater.* **29**, 1905266 (2019).

20. Jiang, C., de Rijk, S. R., Malliaras, G. G. & Bance, M. L. Electrochemical impedance spectroscopy of human cochleas for modeling cochlear implant electrical stimulus spread. *APL Mater.* **8**, 091102 (2020).

21. Moser, M. et al. Side chain redistribution as a strategy to boost organic electrochemical transistor performance and stability. *Adv. Mater.* **32**, e2002748 (2020).

22. Kleemann, H., Krechan, K., Fischer, A. & Leo, K. A review of vertical organic transistors. *Adv. Funct. Mater.* **30**, 1907113 (2020).

23. Giovannitti, A. et al. Energetic control of redox-active polymers toward safe organic bioelectronic materials. *Adv. Mater.* **32**, e1908047 (2020).

24. Wu, X. et al. Enhancing the electrochemical doping efficiency in diketopyrrolopyrrole-based polymer for organic electrochemical transistors. *Adv. Electron. Mater.* **7**, 2000701 (2020).

25. Bischak, C. G., Flagg, L. Q., Yan, K., Li, C. Z. & Ginger, D. S. Fullerene active layers for n-type organic electrochemical transistors. *ACS Appl. Mater. Interfaces* **11**, 28138–28144 (2019).

26. Yan, Y. et al. High-performance organic electrochemical transistors with nanoscale channel length and their application to artificial synapse. *ACS Appl. Mater. Interfaces* **12**, 49915–49925 (2020).

27. Chen, X. et al. n-Type rigid semiconducting polymers bearing oligo(ethylene glycol) side chains for high-performance organic electrochemical transistors. *Angew. Chem. Int. Ed. Engl.* **60**, 9368–9373 (2021).

28. Savva, A. et al. Solvent engineering for high-performance n-type organic electrochemical transistors. *Adv. Electron. Mater.* **5**, 1900249 (2019).

29. Wang, Y. et al. Hybrid alkyl–ethylene glycol side chains enhance substrate adhesion and operational stability in accumulation mode organic electrochemical transistors. *Chem. Mater.* **31**, 9797–9806 (2019).

30. Marks, A. et al. Synthetic nuances to maximize n-type organic electrochemical transistor and thermoelectric performance in fused lactam polymers. *J. Am. Chem. Soc.* **144**, 4642–4656 (2022).

31. Feng, K. et al. Cyano-functionalized n-type polymer with high electron mobility for high-performance organic electrochemical transistors. *Adv. Mater.* **34**, e2201340 (2022).

32. Kim, S. M. et al. Influence of PEDOT:PSS crystallinity and composition on electrochemical transistor performance and long-term stability. *Nat. Commun.* **9**, 3858 (2018).

33. Guo, E. et al. Vertical organic permeable dual-base transistors for logic circuits. *Nat. Commun.* **11**, 4725 (2020).

34. Abarkan, M. et al. Vertical organic electrochemical transistors and electronics for low amplitude micro-organ signals. *Adv. Sci.* **9**, e2105211 (2022).

35. Xie, Z. C. et al. All-solid-state vertical three-terminal n-type organic synaptic devices for neuromorphic computing. *Adv. Funct. Mater.* **32**, 2107314 (2022).

36. Rashid, R. B. et al. Ambipolar inverters based on cofacial vertical organic electrochemical transistor pairs for biosignal amplification. *Sci.*

Adv. **7**, eabh1055 (2021).

37. Lenz, J., Del Giudice, F., Geisenhof, F. R., Winterer, F. & Weitz, R. T. Vertical, electrolyte-gated organic transistors show continuous operation in the MA cm^{-2} regime and artificial synaptic behaviour. *Nat. Nanotechnol.* **14**, 579–585 (2019).

38. Liu, G. et al. Ultralow-Power and multisensory artificial synapse based on electrolyte-gated vertical organic transistors. *Adv. Funct. Mater.* **32**, 2200959 (2022).

39. Paterson, A. F. et al. On the role of contact resistance and electrode modification in organic electrochemical transistors. *Adv. Mater.* **31**, e1902291 (2019).

40. Haddock, J. N. et al. A comprehensive study of short channel effects in organic field-effect transistors. *Org. Electron.* **7**, 45–54 (2006).

41. Thiburce, Q., Giovannitti, A., McCulloch, I. & Campbell, A. J. Nanoscale ion-doped polymer transistors. *Nano Lett.* **19**, 1712–1718 (2019).

42. Venkatraman, V. et al. Subthreshold operation of organic electrochemical transistors for biosignal amplification. *Adv. Sci.* **5**, 1800453 (2018).

43. Lee, S. & Nathan, A. Subthreshold schottky-barrier thin-film transistors with ultralow power and high intrinsic gain. *Science* **354**, 302–304 (2016).

44. Wu, X. et al. Universal spray-deposition process for scalable, high-performance, and stable organic electrochemical transistors. *ACS Appl. Mater. Interfaces* **12**, 20757–20764 (2020).

45. Wu, H. Y. et al. Influence of molecular weight on the organic electrochemical transistor performance of ladder-type conjugated polymers. *Adv. Mater.* **34**, e2106235 (2022).

46. Mannerbro, R., Ranlöf, M., Robinson, N. & Forchheimer, R. Inkjet printed electrochemical organic electronics. *Synth. Met.* **158**, 556–560 (2008).

47. Nilsson, D., Robinson, N., Berggren, M. & Forchheimer, R. Electrochemical logic circuits. *Adv. Mater.* **17**, 353–358 (2005).

48. Doris, S. E., Pierre, A. & Street, R. A. Dynamic and tunable threshold voltage in organic electrochemical transistors. *Adv. Mater.* **30**, e1706757 (2018).

49. Romele, P., Ghittorelli, M., Kovacs-Vajna, Z. M. & Torricelli, F. Ion buffering and interface charge enable high performance electronics with organic electrochemical transistors. *Nat. Commun.* **10**, 3044 (2019).

50. Leydecker, T., Wang, Z. M., Torricelli, F. & Orgiu, E. Organic-based inverters: basic concepts, materials, novel architectures and applications. *Chem. Soc. Rev.* **49**, 7627–7670 (2020).

51. Tzirakis, M. D., Alberti, M. N., Weissman, H., Rybtchinski, B. & Diederich, F. Enantiopure laterally functionalized alleno-acetylenic macrocycles: synthesis, chiroptical properties, and self-assembly in aqueous media. *Chemistry* **20**, 16070–16073 (2014).

52. Chen, J. et al. Tuning the central fused ring and terminal units to improve the photovoltaic performance of Ar(A–D)2 type small molecules in solution-processed organic solar cells. *J. Mater. Chem. A* **4**, 4952–4961 (2016).

53. Song, C. K., Eckstein, B. J., Tam, T. L., Trahey, L. & Marks, T. J. Conjugated polymer energy level shifts in lithium-ion battery electrolytes. *ACS Appl. Mater. Interfaces* **6**, 19347–19354 (2014).

54. Wang, Z. et al. Cinnamate-functionalized natural carbohydrates as photopatternable gate dielectrics for organic transistors. *Chem. Mater.* **31**, 7608–7617 (2019).

55. Wang, K. & Wang, M. Hyperbranched narrow-bandgap DPP homopolymers synthesized via direct arylation polycondensation. *J. Polym. Sci. A Polym. Chem.* **55**, 1040–1047 (2017).

56. Punzi, A. et al. Synthetic Routes to TEG-substituted diketopyrrolopyrrole-based low band-gap polymers. *Eur. J. Org. Chem.* **2016**, 3233–3242 (2016).

57. Izunobi, J. U. & Higginbotham, C. L. Polymer molecular weight analysis by 1H NMR spectroscopy. *J. Chem. Educ.* **88**, 1098–1104 (2011).

58. Zheng, Y. Q. et al. Monolithic optical microlithography of high-density elastic circuits. *Science* **373**, 88–94 (2021).

59. Inal, S., Malliaras, G. G. & Rivnay, J. Benchmarking organic mixed conductors for transistors. *Nat. Commun.* **8**, 1767 (2017).

Acknowledgements

We gratefully acknowledge financial support from the AFOSR (grant nos. FA9550-18-1-0320 and FA9550-22-1-0423), the Northwestern University MRSEC (grant no. NSF DMR-1720139), the US Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD) (award no. 70NANB19H005), the National Natural Science Foundation of China (grant nos. U1830207, 21774055 and 62273073), the National Key R&D Program of China (grant no. 2022YFE0134800), the Sichuan Science and Technology Program (grant no. 2022NSFSC0877) and Flexterra Corp. This work made use of the Northwestern University Micro/Nano Fabrication Facility (NUFAB), the EPIC facility, the Keck-II facility and the SPID facility of the NUANCE Center at Northwestern University, which is partially supported by the Soft and Hybrid Nanotechnology Experimental (SHyNE) Resource (NSF ECCS-2025633), the Materials Research Science and Engineering Center (DMR-1720139), the State of Illinois and Northwestern University. R.M.P acknowledges support from the Intelligence Community Postdoctoral Research Fellowship Program at Northwestern University administered by

Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence (ODNI). N.R.G. and D.M. acknowledge support from AFOSR under grant number FA9550-23RXCOR011. We thank J. Strzalka of the Argonne National Laboratory Advanced Photon Source for assistance with the GIWAXS measurements. Use of the Advanced Photon Source, an Office of Science User Facility operated for the US DOE Office of Science by Argonne National Laboratory, was supported by the US DOE under contract no. DE-AC02-06CH11357. We thank W. Yue and Y. Wang from Sun Yat-Sen University for providing the PIBET-AO polymer. We also thank Z. Ye and C. Wu from Zhejiang University for their discussion and help. W.H. thanks the UESTC Excellent Young Scholar Project for financial support.

Author information

Author notes

1. These authors contributed equally: Wei Huang, Jianhua Chen, Yao Yao

Authors and Affiliations

1. School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China

   Wei Huang, Miao Xie, Libing Bai & Yuhua Cheng

2. Department of Chemistry and the Materials Research Center, Northwestern University, Evanston, IL, USA

   Wei Huang, Jianhua Chen, Yao Yao, Ding Zheng, Liang-Wen Feng, Yao Chen, Robert M. Pankow, Zhi Wang, Tobin J. Marks & Antonio Facchetti

3. Department of Chemical Science and Technology, Yunnan University, Kunming, China

   Jianhua Chen

4. Department of Materials Science and Engineering and the Shenzhen Key Laboratory for Printed Organic Electronics, Southern University of Science and Technology (SUSTech), Shenzhen, China

   Jianhua Chen & Xugang Guo

5. School of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, China

   Yao Yao & Jianfeng Ping

6. Innovation Platform of Micro/Nano Technology for Biosensing, ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China

   Yao Yao & Jianfeng Ping

7. Department of Biomedical Engineering, Northwestern University, Evanston, IL, USA

   Xudong Ji, Abhijith Surendran & Jonathan Rivnay

8. College of Chemistry, Sichuan University, Chengdu, China

   Liang-Wen Feng

9. Air Force Research Laboratory, Materials and Manufacturing DirectorateWPAFB, Ohio, OH, USA

   David Moore & Nicholas R. Glavin

10. School of Materials Science and Engineering, North University of China, Taiyuan, China

    Zhi Wang

11. Flexterra Inc. 8025 Lamon Avenue, Skokie, IL, USA

    Yu Xia & Antonio Facchetti

Contributions

W.H., Y. Cheng, T.J.M. and A.F. conceived the idea and designed the experiments. J.C., R.M.P. and X.G. synthesized the polymer semiconductors. W.H., Y.Y., D.Z., M.X., Y.X., L.B. and J.P. fabricated and characterized the devices. X.J., D.M., N.R.G., A.S. and J.R. conducted the bandwidth and EIS measurements. Y. Chen measured and analysed the GIWAXS. L.-W.F. and Z.W. synthesized the Cin-Cell.

Corresponding authors

Correspondence to Wei Huang, Ding Zheng, Yuhua Cheng, Tobin J. Marks or Antonio Facchetti.

Ethics declarations

Competing interests

A patent application has been filed by Northwestern/Flexterra with inventors W.H., J.C., Y.X., A.F. and T.J.M.

Peer review

Peer review information

Nature thanks Magnus Berggren and Aristide Gumyusenge for their contribution to the peer review of this work.

Additional information

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Synthesis and characterizations of gDPP-g2T and Homo-gDPP.

a, Synthetic route to the semiconductors gDPP-g2T (p-type) and Homo-gDPP (n-type). **b**, ^{1}H NMR spectrum of compound **3**. **c**, ^{13}C NMR spectrum of compound **3**. **d**, ^{1}H NMR spectrum of compound **4**. **e**, ^{13}C NMR spectrum of compound **4**. **f**, ^{1}H NMR spectrum of polymer gDPP-g2T. **g**, ^{1}H NMR spectrum of polymer Homo-gDPP.

Source data

Extended Data Fig. 2 Fabrication process for conventional planar OECTs, performance evaluation for vOECTs based on polymers DPP-2T, Homo-DPP, and vOECTs with and without different Cin-Cell contents in the active layer.

a, Fabrication process for a conventional OECT (cOECT). i) Source/drain Au electrode fabrication (this involves photolithography, Au thermal evaporation, and lift-off processes); ii) spin-coating of the semiconducting layer and UV cross-linking; iii) PBS and Ag/AgCl as electrolyte dielectric and gate electrode. **b**, Chemical structures of the redox-inactive semiconducting polymers **DPP-2T** (p-type); **Homo-DPP** (n-type). **c**, Representative transfer characteristics of p-type **DPP-2T** and n-type **Homo-gDPP** based vOECTs. **d**, Representative transfer characteristics of (**d**) n-type vOECTs with a pristine **Homo-gDPP** and **Homo-gDPP:Cin-Cell** (9:2) as the semiconductor channel and (**e**) p-type vOECTs with a pristine **gDPP-g2T** and **gDPP-g2T:Cin-Cell** (9:2) as the semiconductor channel. W= d = 30 µm, $V_D = \pm 0.1$ V. Representative transfer characteristics of (**f**) **Homo-gDPP:Cin-Cell** and (**g**) **gDPP-g2T:Cin-Cell** based vOECTs with weight ratios of 9:2, 8:3 and 6:4, respectively. W= d = 50 µm.

Source data

Extended Data Fig. 3 Semiconductor film morphologies and microstructures.

a, AFM height images and phase images of a pristine **gDPP-g2T** film and **gDPP-g2T:Cin-Cell blend** (9:2 in mass) film. **b**, AFM height images and phase images of a pristine **Homo-gDPP** film and a **Homo-gDPP:Cin-Cell** blend (9:2 in mass) film. **c**, 2D-GIWAXS images and corresponding (**d**) in-plane and (**e**) out-of-plane one-dimensional line-cuts for **gDPP-g2T**, **gDPP-g2T**: **Cin-Cell**, **Homo-gDPP**, **Homo-gDPP**: **Cin-Cell**, and **Cin-Cell** films. The **gDPP-g2T** in both pristine and blend films exhibits a preferential π-face-on orientation, yet with a considerable portion of crystallites being edge-on oriented. In all cases the lamellar and π-π periodicities are pinned at 20.2 Å and 3.6 Å, respectively. However, in both **Homo-gDPP** and **Homo-gDPP:Cin-Cell** the polymer orientation is preferentially edge-on and while the π-π distance is pinned at 3.6 Å, and the lamellar spacings vary slightly, from 20.2 Å to 19.6 Å, respectively.

Source data

Extended Data Fig. 4 OECT performance.

(**a**,**c**) Representative transfer characteristics and (**b**,**d**) corresponding g_m and SS curves of (**a**,**b**) p-type **gDPP-g2T:Cin-Cell** and (**c**,**d**) n-type **Homo-gDPP:Cin-Cell** vOECTs with the indicated W and d values ranging from 30, 50, and 70 µm. vOECT performance with unpatterned semiconducting layers, where **e** and **f** are representative transfer characteristics and the corresponding g_m and SS curves of an unpatterned p-type **gDPP-g2T:Cin-Cell** vOECT (W= d = 30 µm) and an unpatterned n-type **Homo-gDPP:Cin-Cell** vOECT (W= d = 50 µm), respectively. Impressively, these devices exhibit almost identical performance parameters (I_{ON}, V_{ON}, g_m) versus those of the corresponding p- and n-type vOECTs having a patterned channel, with the exception of a higher off-current for the p-type vOECT. This is due to confinement of the charge transport within the overlapping source/drain electrodes, and the ultra-small L prevents substantial fringing effects. Representative transfer characteristics and corresponding g_m and SS curves for (**g**) p-type **gDPP-g2T:Cin-Cell** cOECTs and (**h**) n-type **Homo-gDPP:Cin-Cell**. Note, W = 100 µm, L = 10 µm.

Source data

Extended Data Fig. 5 Chemical structures and OECT performance of the indicated polymer semiconductors blended with Cin-Cell.

a, Chemical structures of the redox-active semiconducting polymers [Pg2T-T and PIBET-AO (p-type); PEG-N2200 and BTI2 (n-type)]. Representative transfer and g_m characteristics of planar **(b**, **d**, **f**, **and h**) and vertical (**c**, **e**, **g**, **and i**) OECTs based on Pg2T-T (**b**,**c**), PIBET-AO (**d**,**e**), PEG-N2200 (**f**,**g**), and BTI2 (**h**,**i**). Note, W and d are both 30 µm for the vOECT, while W and L are 100 µm and 10 µm, respectively, for the cOECT. For PEG-N2200 and BTI2 based planar OECTs, the g_ms are too small when plotted on the same scale as that of the vOECT counterparts, so therefore insets with magnified Y axis are included. In all samples, the semiconductor:Cin-Cell weight ratio is 9:2.

Source data

Extended Data Fig. 6 Vertical device structures for the direct observation of the electrochromic and performance characteristics of vOECTs with different top electrode widths.

a, Top view photograph and (**b**) corresponding p-type vOECT schematic indicating the large electrode overlapping area (~2 × 0.5 mm) used to monitor the electrochromic process associated to the redox chemistry of semiconducting layer. Here a L-shaped bottom Au electrode (100 nm thick) and an L-shaped top Au electrode (20 nm) are grounded and biased with −0.1 V, respectively. **c**, Cross section illustration of the vOECT from the indicated red dashed line in (**b**). **d**, Optical image of a p-type gDPP-g2T based vOECT array with different top electrode widths but identical bottom electrode. Note, these devices were fabricated by photolithography to accurately pattern the electrode dimensions and locate the semiconductor at the electrode cross points. **e**, Representative transfer characteristics of p-type and n-type (Homo-gDPP) vOECTs for the indicated top electrode width. **f**, Dependence of g_m on top electrode width for the p- and n-type vOECTs, where V_D = −0.5 V and 0.5 V, respectively. Data are from the average of 8 different devices. **g**, Transfer and g_m characteristics of PEDOT:PSS based

vOECTs (W = d = 30 μm) and cOECTs (W = 100 μm, L = 10 μm);, V_D = −0.1 V.

Source data

Extended Data Fig. 7 Transfer characteristics, output characteristics, transient properties, and bandwidth characteristics of OECTs.

Transfer characteristics of (**a**) n-type **Homo-gDPP:Cin-Cell** and (**b**) p-type **gDPP-g2T:Cin-Cell** vOECTs with V_D from ±1 mV to ±500 mV. W = d = 30 μm. Output characteristics of (**c**) n-type **Homo-gDPP:Cin-Cell** and (**d**) p-type **gDPP-g2T:Cin-Cell** vOECTs with W = d = 30, 50, and 70 μm. Cycling stability (cycling frequency of 10 Hz) of PEDOT:PSS based (**e**) cOECT and (**f**) vOECT, where V_D = −0.1 V, V_G is switching between −0.1 V and 0.7 V with a frequency of 10 Hz. **g**, Illustration of the cOECT and vOECT structures characterized by transient and bandwidth measurements. **h**, Drain current transient responses of cOECTs based on **gDPP-g2T:Cin-Cell**(9:2) and **Homo-gDPP:Cin-Cell**(9:2). (Semiconductor thicknesses are 100 nm, V_G is a square pulse from 0.0 to −0.5 V and 0.0 to 0.7 V for p-type and n-type cOECTs, respectively). **i**, Frequency dependent transconductance of cOECTs based on **gDPP-g2T:Cin-Cell** (9:2), and **Homo-gDPP:Cin-Cell** (9:2) (Error bars represent s. d. for N = 6). Cutoff frequencies (f_c) of ~800 and ~500 Hz were measured for **gDPP-g2T** based cOECT and vOECT, respectively. This result is consistent with the transient response, where both τ_{ON} and τ_{OFF} of the vOECT are greater than those of the corresponding cOECT. In contrast, the f_c of Homo-gDPP based vOECT is ~1200 Hz, which is much higher than that in the corresponding cOECT (~ 600 Hz). This reflects the unstable cycling behavior of Homo-gDPP based cOECTs, which degrades the transconductance during bandwidth measurements, especially at high frequencies. Note, in the vertical structure, the cycling stability is greatly enhanced due to the protection of the top electrode which prevents doped polymer film dissolution/delamination, thereby maintaining structural stability.

Source data

Extended Data Fig. 8 Electrochemical impedance spectroscopy, transient response, and ion sensing using vOECT-based vertically stacked inverters.

a, EIS measurement setup of typical two-electrode along with the corresponding spectroscopies based on (**b**) **gDPP-g2T:Cin-Cell**(9:2) and (**c**) **Homo-gDPP:Cin-Cell** (9:2). **d**, Drain current transient responses of vOECTs with **gDPP-g2T:Cin-Cell** (9:2) thicknesses of 100 nm, 200 nm, 300 nm, 400 nm. V_G is a square pulse between 0.0 and −0.5 V. **e**, Voltage output characteristics of the vOECT-based vertically stacked complementary inverter for ion sensing over a wide concentration range of KCl in water. **f**, Measured transition voltage as a function of KCl concentration. Sensitivities of 31 mV/dec ($10^{-6} \sim 10^{-3}$ M) and 89 mV/dec ($10^{-3} \sim 1$ M) are obtained with linear approximation (dash line) of the measurements.

Source data

Extended Data Fig. 9 Fabrication process for the Vertically stacked complementary inverter (VSCI) and circuits.

a, *For the Inverter:* i). Bottom electrodes (V_{DD}) evaporation; ii). P-type semiconducting layer fabrication; iii). Middle electrode (V_{OUT}) evaporation; iv). N-type semiconducting layer fabrication; v). Top electrodes (GND) evaporation; vi). Application of the PBS electrolyte and Ag/AgCl electrode. **b**, *For the Ring Oscillator* (enlarged plots are also provided)*:* i). Bottom Au electrodes (V_{DD}) evaporation; ii). P-type semiconducting layer fabrication; iii). Middle electrodes (V_{OUT}) evaporation; iv). N-type semiconducting layer fabrication; v). Top electrodes (GND) evaporation; vi). Cin-Cell encapsulation layer fabrication, where the channel areas and Middle electrode acting as V_{OUT} are left open; vii). Application of the Ag/AgCl paste on top of the Middle electrode V_{OUT} region; viii) Apply the PBS electrolyte. **c**, *For the NAND:* i). Bottom electrodes (V_{DD}, and GND) evaporation; ii). P-type semiconducting layer fabrication; iii). Middle electrodes (V_{OUT}) evaporation; iv). N-type semiconducting layer fabrication; v). Top electrodes fabrication; vi). Cin-Cell encapsulation layer fabrication,

where the channel areas are left open; vii). Apply the PBS electrolyte and Ag/AgCl electrode. **d**, *For the NOR:* i). Bottom electrodes evaporation; ii). P-type semiconducting layer fabrication; iii). Middle electrodes (V_{OUT}) evaporation; iv). N-type semiconducting layer fabrication; v). Top electrodes (V_{DD}, and GND) evaporation; vi). Cin-Cell encapsulation layer fabrication, where the channel areas are left open; vii). Apply the PBS electrolyte and Ag/AgCl electrode. **e**, *For the rectifier*: i). Bottom Au electrodes (V_{OUT}) evaporation; ii). P-type semiconducting layer fabrication; iii). Middle electrodes (V_{IN+}, and V_{IN-}) evaporation; iv). N-type semiconducting layer fabrication; v). Top electrodes (GND) evaporation; vi). Cin-Cell encapsulation layer fabrication, where the channel areas and the middle electrode acting as V_{IN} are left open; vii). Application of the Ag/AgCl paste on top of the middle electrode V_{IN} region; viii). Apply the PBS electrolyte.

Extended Data Table 1 Summary of OECT performance metrics

Supplementary information

Supplementary Video 1

Top view video of a p-type gDPP-g2T:Cin-Cell vOECT with a large electrode overlapping area (approximately 2 × 0.5 mm^2). The electrochromic process associated with the redox chemistry of the semiconducting layer can be observed in this video. Here the bottom Au electrode (100 nm) and the top Au electrode (20 nm) are biased with 0 and −0.1 V, respectively. The voltage bias applied on Ag/AgCl electrode is indicated in the video.

Supplementary Video 2

Top view video of an electrolyte capacitor based on a p-type gDPP-g2T:Cin Cell film. The voltage bias applied on the Ag/AgCl electrode is indicated in the video. Note that the ellipsoid visible in the middle of the video is the reflection of the microscope light illuminating the device during the recording. When V_G is switched from 0 V to −0.7 V, the polymer film located in direct contact with the Au electrode oxidizes immediately (in 100

ms in a 2 × 0.5 mm^2 area). This reflects the high electric field (E) near Au increasing the ion drift velocity, s ($s = \mu_i E$, where μ_i = ion mobility), thus promoting faster doping. Next, as the charging process continues, the entire electrolyte area covered semiconductor (approximately 6 × 2.5 mm^2) oxidizes slowly (approximately 2 s), starting from the semiconductor near the Au electrode. Similarly, the reduction also begins from the semiconductor in direct contact with the Au electrode, followed by the portion near the edges. Therefore, since the entire vOECT channel layer is in direct contact with the Au electrode, the redox process is intrinsically fast.

Source data

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 4

Source Data Extended Data Fig. 1

Source Data Extended Data Fig. 2

Source Data Extended Data Fig. 3

Source Data Extended Data Fig. 4

Source Data Extended Data Fig. 5

Source Data Extended Data Fig. 6

Source Data Extended Data Fig. 7

Source Data Extended Data Fig. 8

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Huang, W., Chen, J., Yao, Y. *et al.* Vertical organic electrochemical transistors for complementary circuits. *Nature* **613**, 496–502 (2023). https://doi.org/10.1038/s41586-022-05592-2

- Received: 05 July 2021
- Accepted: 24 November 2022
- Published: 18 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05592-2

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

Vertical architecture improves performance of transistor family

- Camille Cunin
- Aristide Gumyusenge

Nature News & Views 18 Jan 2023

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05592-2

- Article
- Open Access
- Published: 18 January 2023

Modern temperatures in central–north Greenland warmest in past millennium

- M. Hörhold ,
- T. Münch,
- S. Weißbach,
- S. Kipfstuhl,
- J. Freitag,
- I. Sasgen,
- G. Lohmann,
- B. Vinther &
- …
- T. Laepple

Show authors

Nature **volume 613**, pages 503–507 (2023)

- 9081 Accesses
- 1011 Altmetric
- Metrics details

Abstract

The Greenland Ice Sheet has a central role in the global climate system owing to its size, radiative effects and freshwater storage, and as a potential tipping point[1]. Weather stations show that the coastal regions are warming[2], but the imprint of global warming in the central part of the ice sheet is unclear, owing to missing long-term observations. Current ice-core-based temperature reconstructions[3,4,5] are ambiguous with respect to isolating global warming signatures from natural variability, because they are too noisy and do not include the most recent decades. By systematically redrilling ice cores, we created a high-quality reconstruction of central and north Greenland temperatures from ad 1000 until 2011. Here we show that the warming in the recent reconstructed decade exceeds the range of the pre-industrial temperature variability in the past millennium with virtual certainty ($P < 0.001$) and is on average 1.5 ± 0.4 degrees Celsius (1 standard error) warmer than the twentieth century. Our findings suggest that these exceptional temperatures arise from the superposition of natural variability with a long-term warming trend, apparent since ad 1800. The disproportionate warming is accompanied by enhanced Greenland meltwater run-off, implying that anthropogenic influence has also arrived in central and north Greenland, which might further accelerate the overall Greenland mass loss.

Main

Global mean temperature has increased to 1 °C above pre-industrial levels in the second decade of the twenty-first century[6]. Regionally, the Arctic shows the strongest warming[7], particularly in winter[8]. However, to quantify how extraordinary the recently observed temperature changes are, they have to be placed in the context of past temperatures and natural climate variability. For this, instrumental records are often too short, and although climate models are able to reproduce long-term trends[9], they tend to underestimate regional climate variability[10,11] and are challenging to validate. Thus, temperature reconstructions from palaeoclimate proxies are essential for estimating pre-industrial natural climate variability. However, most large-scale reconstructions that are based on multiple proxy types or tree ring records require a proxy screening and instrumental calibration step and thus might

be prone to underestimation of past climate variability outside of the calibration period[12].

For the Arctic, the regional temperature reconstruction Arctic 2k[13] shows a persisting warming trend since the nineteenth century and the emergence of air temperature values outside the natural (pre-industrial) variability since the early-mid twentieth century[14]. Increasing temperatures in the Arctic also affect the Greenland Ice Sheet, causing more mass loss by increasing meltwater run-off[1,15]. Weather stations from the Greenland coast cover 200 years[16] and indicate delayed warming trends compared to other regions[17,18] with large regional and seasonal differences[2] along the coast at the beginning of the twenty-first century. Although the melt area of the ice sheet has been observed to progress towards higher elevations[19], little is known about the magnitude and the trend of the surface temperature changes in the central parts of the ice sheet. The reasons lie in the short instrumental records, as well as the sparsity of palaeoclimate data, their low spatial or temporal coverage, and the high noise level in the records.

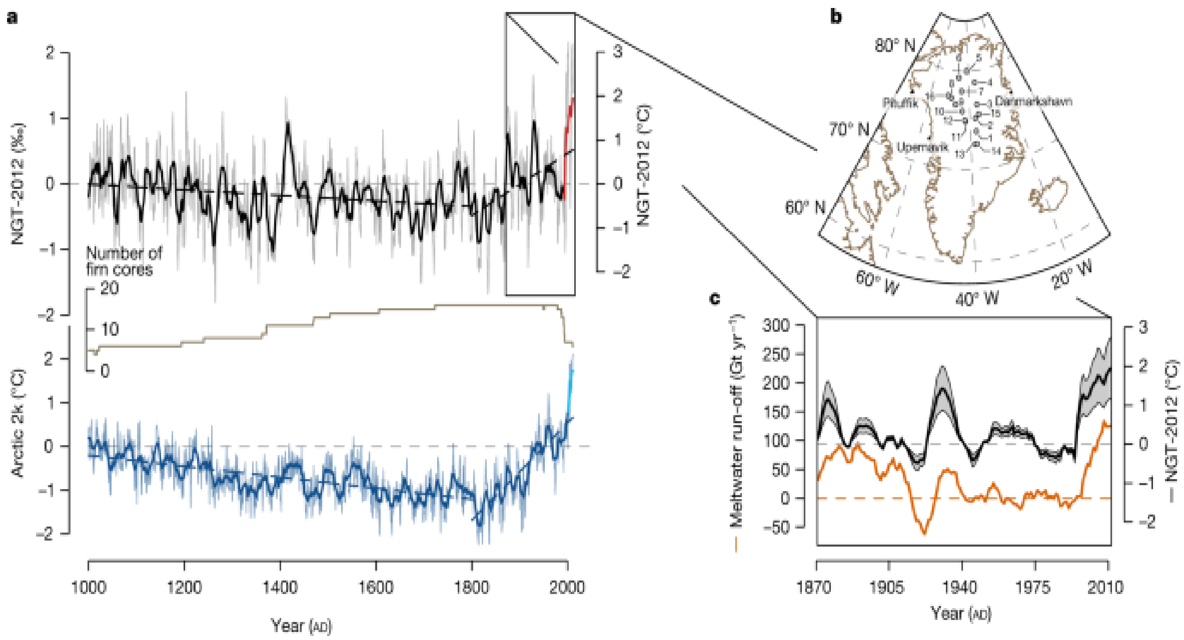
Previous ice-core data from central and north Greenland provide an inconclusive picture of the imprint of anthropogenic forcing on the surface temperature, either owing to short temporal coverage[3,20], or, because the records are based on single ice-core sites[4], owing to uncertainty on the strength and representability of the contained climate signal[21]. The only available multisite stacked climate record, originating from the North Greenland Traverse (NGT), did not indicate signatures of warming but ended in ad 1995 (ref. [5]).

The NGT-2012 record

To analyse the Greenland temperature evolution over the past decades with respect to natural variability and global warming, we here extend the previous NGT reconstruction to the year 2011 (all dates are ad). In 2011 and 2012, five of the NGT ice-core sites were revisited and shallow firn cores were taken near the original drilling sites to complement the existing records (Methods). Altogether, the new record (‘NGT-2012’, Fig. 1a) is stacked from a compilation of 21 stable oxygen isotope records ($\delta^{18}O$ anomalies

relative to the 1961–1990 reference interval; Methods) from north and central Greenland (Extended Data Table 1). NGT-2012 covers more than 1,000 years, providing unprecedented spatial and temporal coverage of the area (Fig. 1b). Single records of stable isotope data exhibit a large proportion of non-climatic noise[21]. Here we combine many records to improve the signal-to-noise ratio of our reconstruction (Extended Data Fig. 1). A comparison between individual time series constituting the NGT-2012 stack shows a substantial spatially coherent signal on decadal and longer timescales with a signal-to-noise ratio greater than 3 (Extended Data Fig. 1). We hence apply an 11-year running mean filter to our time series to focus our analyses on these timescales (‘decadal temperatures’).

Fig. 1: The millennial NGT-2012 temperature reconstruction record from Greenland.



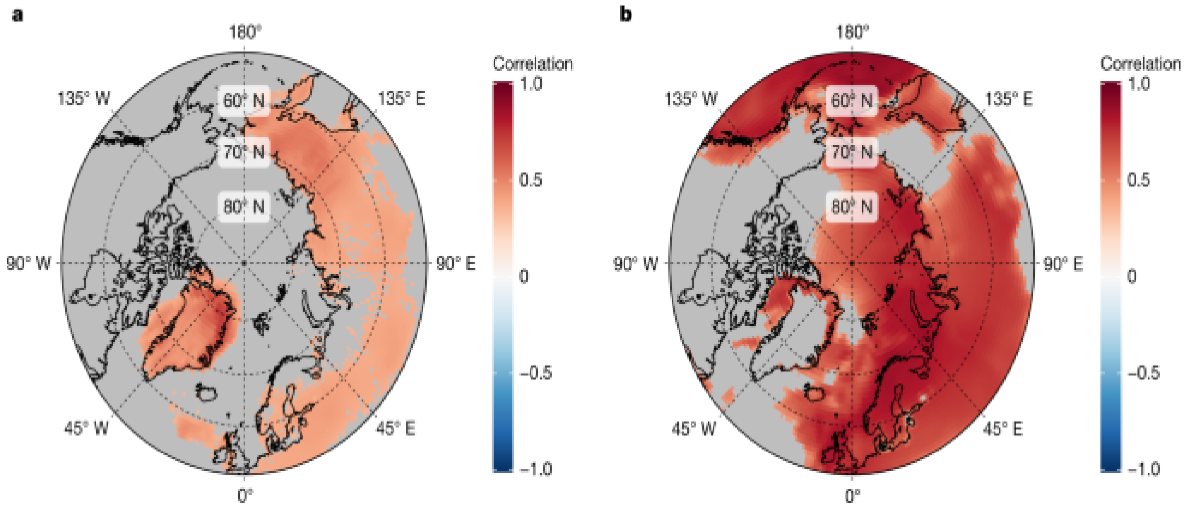
a, The NGT-2012 composite record of 11-year running mean $\delta^{18}O$ (black, left axis) and inferred temperature time series (right axis, Methods) from 1000 to 2011 (top panel). Light grey lines in the background display annual mean values. The thick red line highlights the extension of existing ice-core records to 2011 by re-drillings performed as part of this study. Estimated linear trends over the periods 1000–1800 (pre-industrial) and 1800–2011 are shown as dashed black lines. The number of firn cores contributing to the reconstruction is shown beneath as a brown line. The bottom panel shows the Arctic 2k temperature reconstruction record[13] displayed as 11-year

running mean values and as annual data and with dashed blue lines indicating linear trends, as for NGT-2012. The time series was extended to 2011 using HadCrut instrumental data[70] (cyan line, Methods). **b**, Locations of the ice cores used for NGT-2012 (circles) and of nearby weather stations[16] (black triangles; geographic map data obtained from the 'rnaturalearth' package for the software R). Site IDs are detailed in Extended Data Table 1. **c**, Comparison of the NGT-2012 11-year running mean temperature reconstruction (1871–2011, black) with Greenland meltwater run-off from MAR3.5.2[22] ($R = 0.62$, $P < 0.01$, $n = 141$; Methods and Extended Data Fig. 7). Grey shading indicates a ±40% uncertainty of the temperature reconstruction obtained from the range of plausible calibration slopes (Methods). All time series are displayed as anomalies relative to the 1961–1990 reference period (horizontal dashed lines).

Source data

The NGT-2012 stack exhibits a strong correlation ($R \geq 0.75$, $P < 0.01$, $n = 111$) with the decadal annual mean air temperatures from weather stations along the Greenlandic coast[16] over the common period (Methods and Extended Data Fig. 2a). In addition, we find a correlation of $R = 0.76$ ($P < 0.001$, $n = 141$) with the decadal annual mean air temperatures at the coring sites derived from the regional climate model MAR3.5.2[22] (1871–2011; Extended Data Fig. 2b; Methods). This relationship also holds true when comparing NGT-2012 to near-surface air temperatures from the Twentieth Century Reanalysis dataset (20CRv3)[23,24] across the region of the NGT-2012 ice-core locations (20CR@NGT; $R = 0.62$, $P < 0.005$, $n = 176$; Methods and Fig. 2a). Together, this shows that the stack can be safely interpreted as a spatially representative temperature record for central and north Greenland over the past millennium.

Fig. 2: NGT-2012 and Arctic 2k[13] point correlation with the 20CRv3[23,24] near-surface temperature field.



a, Point correlation between the 20CRv3[23,24] reanalysis field of 11-year running mean near-surface temperature and the NGT-2012 11-year running mean $\delta^{18}O$ temperature reconstruction time series. **b**, As in **a** but for the point correlation with the Arctic 2k[13] 11-year running mean temperature reconstruction time series. Correlations are calculated for the time period 1836–2000 for all reanalysis grid cells ≥50° N. Grid cells filled grey mark areas with non-significant correlation values ($P > 0.05$, $n = 165$). All geographic map data are obtained from the 'rnaturalearth' package for the software R.

Source data

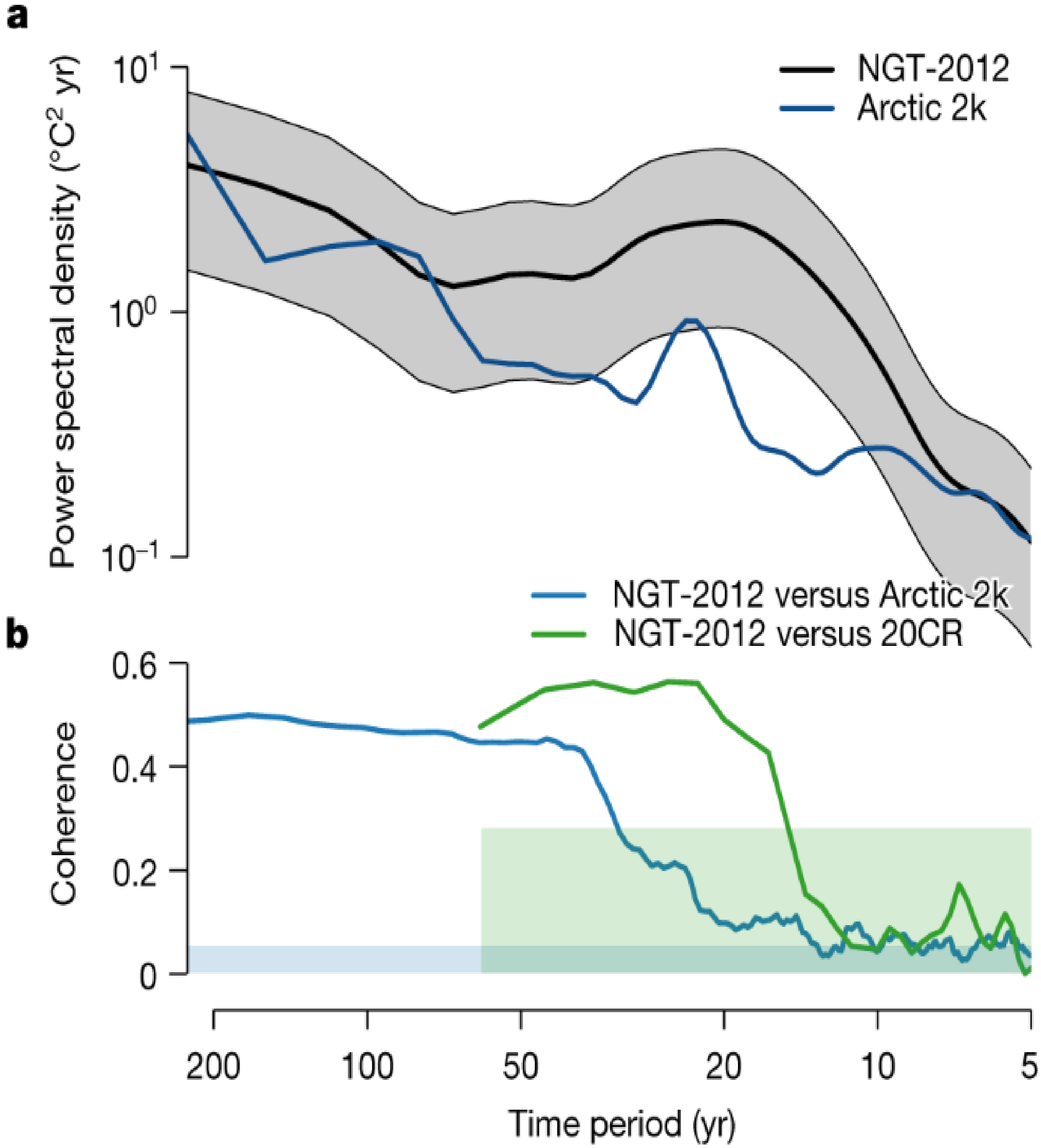
To estimate temperature anomalies, we apply the spatial calibration slope for Greenland of 1/0.67 °C per ‰ (ref. [25]) and use the range of published slopes as an uncertainty (Methods and Fig. 1c). This straightforward approach avoids biases from temporal calibrations or screening against the instrumental record that affect commonly used reconstructions[12].

Natural variability and recent warming

The NGT-2012 temperature record shows a cooling trend from 1000 towards 1800 (−0.06 ± 0.01 °C per 100 years; ±1 standard deviation), followed by a warming trend until 2011 (0.70 ± 0.11 °C per 100 years; Fig. 1a). To characterize the natural climate variability on the Greenland Ice Sheet we analyse the NGT-2012 temperature record using power spectral analysis

including a noise correction for the ice-core stack (Methods). We find a broad maximum in the spectral power of the NGT-2012 temperature for time periods from 11 to 51 years, indicating pronounced natural variability at decadal to bi-decadal timescales (Fig. 3a).

Fig. 3: NGT-2012 and Arctic 2k[13] temperature power spectra and coherence.

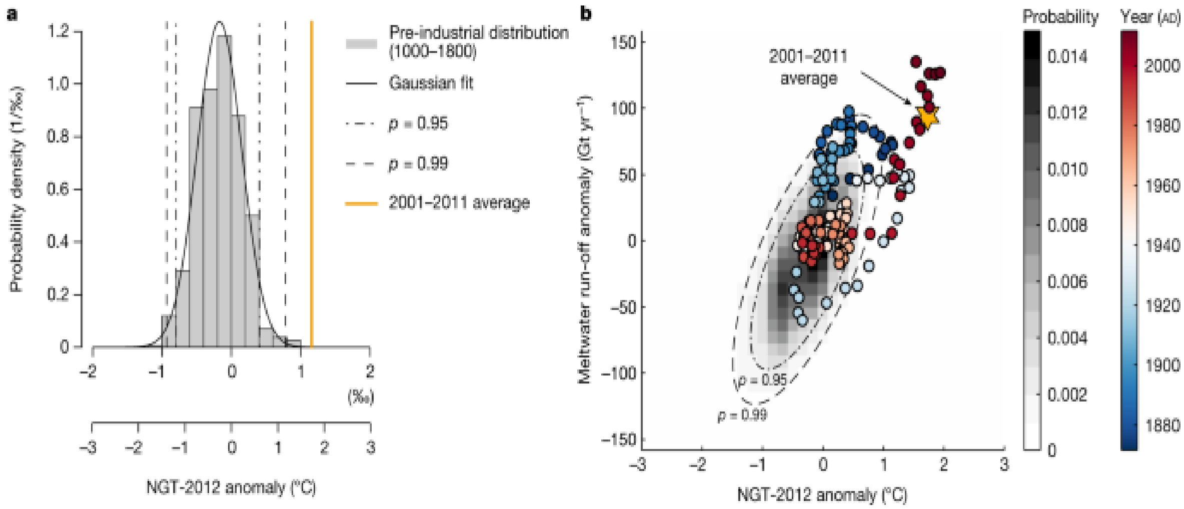


a, The NGT-2012 spectrum (black) represents the signal content for the 1505–1978 time period, which is common to all individual ice cores from the stacked record after removing the local noise contribution (Methods); grey shading denotes the spectral uncertainty range obtained by applying different plausible temperature calibrations (Methods). The Arctic 2k[13] spectrum (dark blue) shows the power spectral density of the 1000–2011 time series. Notably, the average power of the variability in the timescale range from 11 to 51 years is 1.5 to 8 times larger for NGT-2012 than for Arctic 2k (depending on the temperature calibration). **b**, The magnitude-squared coherence of NGT-2012 and Arctic 2k (blue) and of NGT-2012 and 20CRv3[23,24] averaged across the region of the NGT-2012 cores (20CR@NGT, green). Blue and green shadings indicate the respective local 95% confidence level for the coherence based on surrogate data (Methods).

Source data

The reconstructed temperature of the 2001–2011 decade is found to be on average 1.7 ± 0.4 °C (±1 standard error) warmer than the 1961–1990 reference interval and 1.5 ± 0.4 °C warmer than the twentieth century (Methods). Despite the pronounced natural variability that we observe, this high temperature value is exceptional in the context of the past 1,000 years. The 2001–2011 decadal average of the temperature anomalies lies clearly outside the distribution of the pre-industrial values of 1000–1800 (Fig. 4a), with a likelihood for the recent value to occur under the pre-industrial distribution close to zero ($P = 1.82 \times 10^{-5}$, Methods). This result is robust against different variations of creating and analysing the NGT-2012 stack (Extended Data Fig. 3 and Extended Data Table 2), and holds true for timescales shorter than decadal (running mean filter windows <11 years; $P \approx 10^{-4}$; Extended Data Table 2). The recent extreme in temperature can thus be considered as superposition of the anthropogenic global warming trend and pronounced natural variability[26,27], which can also explain the ambiguous signatures of warming, or the lack thereof, in earlier observations from the central and northern Greenland Ice Sheet[3,5,18,27,28].

Fig. 4: Probability density distributions of past Greenland temperature and meltwater run-off.



a, Histogram of the NGT-2012 11-year running mean $\delta^{18}O$ values and of the related reconstructed temperatures (Methods) for the pre-industrial time period (1000–1800) together with a Gaussian fit (thick black line). Vertical dashed lines show the quantiles corresponding to probabilities of $p = 0.95$ and $p = 0.99$, respectively. The 2001–2011 block-averaged NGT-2012 $\delta^{18}O$ and temperature value is shown as a thick orange line. **b**, Meltwater run-off anomaly from MAR3.5.2[22] as a function of the NGT-2012 temperature anomaly over the common time period 1871–2011 (coloured points; the 2001–2011 mean value is indicated as the orange star) together with the reconstructed pre-industrial probability density distribution (grey shading). The pre-industrial meltwater run-off distribution is obtained from a linear regression of MAR3.5.2[22] meltwater run-off against NGT-2012 temperature over the common period (Methods; Extended Data Fig. 2g). Dashed ellipses indicate the area corresponding to probabilities of $p = 0.95$ and 0.99, respectively. We note that the high number of outliers is probably due to the increase of NGT-2012 after 1800 (Fig. 1) that might be related to the early onset of industrial-era warming[14].

Source data

Greenland and Arctic-wide temperatures

The Arctic-wide temperature reconstruction Arctic 2k shows a stronger cooling trend until 1800 as compared to NGT-2012, and a stronger warming trend thereafter (Fig. 1a). Throughout the past millennium, our ice-core-

based Greenland temperature reconstruction and the Arctic-wide temperature reconstruction are correlated ($R = 0.65$, $P < 0.001$, 1000–2011), but this correlation does not persist when limiting the comparison to the twentieth century ($R = 0.28$, $P = 0.17$, $n = 100$; Methods and Extended Data Fig. 4), the time period that arguably has the best reconstruction quality.

To gain further insight into the relationship between NGT-2012 and Arctic 2k, we analyse their spatial representativeness by calculating point correlation maps with the 20CRv3 reanalysis temperature dataset[24] (Fig. 2). This reveals that both reconstructions represent complementary geographic regions. The Arctic 2k reconstruction represents large parts of the higher Arctic circumpolar region but only shows a low correlation over Greenland (Fig. 2b). In a first look, this is surprising because a number of Greenland ice-core records are included in the reconstruction. By contrast, the NGT-2012 record exhibits significant positive regional correlations over the ice sheet (Fig. 2a) and is almost solely representative for Greenland—a result which is also robust for annual mean values that are subject to more reconstruction uncertainty (Extended Data Fig. 5a,b). The distinct spatial correlation structure is not an artefact of the reconstructions. Replacing the NGT-2012 reconstruction by the temperature extracted from the 20CRv3 reanalysis for the region represented by NGT-2012 and Arctic 2k, respectively, results in virtually the same complementary patterns (Extended Data Fig. 5c,d). Mechanisms to explain the weak correlation between the Arctic region and the Greenland Ice Sheet include different elevation[18,28] and thus different changes in wind, cloud cover or radiation pattern over the ice sheet[3] and the distinct effect of circulation variability and changes on Greenland temperature[29,30].

Our results strengthen the observation that the temperature evolution on the Greenland Ice Sheet is partially separated from that of the remaining Arctic. This implies that one single time series alone does not provide a good representation of the Arctic temperature evolution. Here, our Greenland reconstruction and Arctic 2k together provide a more complete picture in the assessment of past and recent temperature changes in the circum-Arctic region and are an important step towards spatio-temporal reconstructions of the Arctic temperature evolution.

The decoupling is visible also in the distinct spectrum of temperature variability. Both temperature reconstructions show pronounced decadal to multidecadal variability (11–51 year time periods; Fig. 3a), which is in agreement with other findings for the Arctic region[26,31,32,33], but for NGT-2012 the variability is four times larger compared to Arctic 2k (range 1.5–8, depending on the temperature calibration). At the same time, both temperature reconstructions exhibit similar power spectral densities for time periods above 50 years and below 8 years. This indicates that the different spatial coverage of the reconstructions is not the primary reason for the variability difference for decadal to multidecadal time periods as a different spatial coverage is expected to mainly affect the short timescales.

Analysing the timescale-dependent relation of the Greenland and Arctic reconstructions shows a high coherence at time periods longer than 50 years, which, however, drops towards shorter time periods (Fig. 3b). By comparison, the coherence between NGT-2012 and the local temperature (NGT@20CRv3) remains high down to time periods below 20 years. This demonstrates that the decoupling between NGT-2012 and Arctic 2k on the decadal to multidecadal timescales is not an artefact of the NGT reconstruction quality.

Thus, the strong temperature variability in the NGT-2012 record probably originates from a regional specific climate signal such as Greenland blocking[29,34,35], an atmospheric variability pattern associated with the negative phase of the North Atlantic Oscillation (NAO). As greater geopotential heights are thermodynamically linked to higher temperatures[30], prolonged atmospheric blocking episodes—that is, persistent high-pressure systems over Greenland—may lead to the northward advection of warm air[36], and accordingly to increased temperatures on the ice sheet[29,36,37]. Indeed, we find a significant correlation ($R = 0.63$, $P < 0.005$, $n = 161$) between the NGT-2012 temperature record and the Greenland Blocking Index (GBI[29]; Extended Data Fig. 2d,e), supporting Greenland blocking as one reason for the larger variability at decadal time scales of the NGT-2012 record compared to Arctic 2k. Greenland blocking was suggested to influence surface melt by influencing the advection of warm air masses[36,37]. In support of this, we find a high correlation between GBI and the Greenland meltwater run-off, derived from the regional climate model MAR3.5.2[22] for

the time period 1871–2011 ($R = 0.80$, $P < 0.001$, $n = 141$; Methods and Extended Data Fig. 2f). During the past decades, the GBI increased in frequency, and to some extent, persistence and magnitude, particularly in summer[37]. This indicates that blocking conditions[34,35,36], superimposed on thermodynamic warming and natural decadal temperature variability, have contributed to the observed records of summer melt in Greenland.

Greenland's future meltwater run-off

Greenland has become a major source of mass-related sea level rise[38,39,40,41] in the past decade, exceeding thermal expansion and contribution from other glaciers, owing to a strong reduction of its surface mass balance by increased summer melt production[42]. In low-elevation areas, the increased surface air temperatures, changes in albedo and the radiation budget, as well as the decreased capacity of meltwater retention in the firn[1,28,38,43,44], have enhanced meltwater run-off. At the same time, the area undergoing summer melt steadily progresses upwards to higher elevations[19,28,45]. For the period 1871–2011 we find a strong connection ($R = 0.62$, $P < 0.01$, $n = 141$; Methods) between the high-elevation NGT-2012 temperature anomaly and meltwater run-off of the ice sheet. These findings emphasize that increased atmospheric temperatures at high elevations in central and north Greenland are indicative of an increased number and intensity of large melt events, probably also in the future[15]. In principle, the higher meltwater run-off could be partly compensated by an increase of accumulation accompanying the warmer temperatures. Whereas accumulation reconstructions from the NGT-2012 stack are much more uncertain than the NGT-2012-based temperature reconstruction (Methods), they do not provide evidence for a strong link of temperature and accumulation or unprecedented accumulation in the past decade (Extended Data Fig. 6).

The strong statistical and physically meaningful relationship between the NGT-2012 record and the meltwater run-off enables us to generate the first reconstruction of the meltwater run-off anomalies for Greenland over the past millennium and thus to put the recent run-off anomalies into the long-term context (Fig. 4b and Methods). The meltwater run-off anomalies of the 2001–2011 decade are outside the reconstructed distribution of pre-industrial (1000–1800) values taking into account the reconstruction uncertainties in

our linear model. Therefore—although with less certainty than for the temperature—our analysis suggests that current decadal meltwater run-off anomalies are unprecedented over the past millennium. This will probably affect the firn densification and the potential for meltwater storage[19,28,40,46] with further implications for the ice sheet mass balance.

In addition to these findings, our meltwater run-off reconstruction provides a baseline to model past and future freshwater discharge from Greenland[47,48] and their effects on the ocean dynamics, for example the Atlantic Meridional Overturning Circulation[49,50].

Our findings demonstrate that recent temperatures in central and north Greenland are higher than in the past 1,000 years and thus demonstrate that global warming is now also detectable in one of the most remote regions in the world. Likewise, meltwater run-off observed today is probably unprecedented over the past millennium. As warming supports an increased frequency of more widespread summer melt events, reaching in some occasions also central and north Greenland, firn properties such as permeability and meltwater retention may change, comparable to firn changes observed in warmer, and lower-elevation areas. Combined with the finding that temperatures in central and north Greenland and meltwater run-off in the ablation zone are already unprecedented compared to the past millennium, an increasing mass loss of the ice sheet is expected under further global warming.

Methods

Dataset

We compiled a set of 21 annually resolved records of relative stable isotopic composition ($\delta^{18}O$; that is, the deviation of the ratio of oxygen-18 to oxygen-16 isotopes in the sample from the respective mean ratio in the global ocean, expressed in per mille, and widely used as a temperature proxy) from central and north Greenland (Extended Data Table 1). For all 21 $\delta^{18}O$ records we use the anomaly time series relative to the 1961–1990 mean value in all further analyses. Five of these records are derived from new shallow firn cores obtained between the years 2011 (B26-2012)[51] and 2012

(B18-2012, B21-2012, B23-2012 and NGRIP-2012) to extend the existing $\delta^{18}O$ records originating from the 1993–1995 North Greenland Traverse[5] and from the location of the North Greenland Ice Core Project (NGRIP) deep ice core[52].

The extension cores were measured in the field for di-electrical profiling using the set-up for the North Greenland Eemian Ice Drilling (NEEM) ice core[53] to derive dating tie points by matching against known volcanic eruptions. The cores B18-2012, B21-2012, B23-2012 and NGRIP-2012 were processed and analysed in the cold room facilities of the Alfred Wegener Institute in Bremerhaven, Germany. Firn density was measured by means of two-dimensional X-ray microtomography[54] with a 0.1-mm resolution and the resulting density profiles were smoothed with a Gaussian filter applying a window size of 2 cm. Stable isotopic composition was measured using cavity ring-down spectrometer instruments (L2120-i and L2130-i, Picarro) following the protocol of a previous work[55]. Measurement uncertainty for $\delta^{18}O$ is smaller than 0.1‰. Dating was performed by annual layer counting based on the isotopic composition and the smoothed density profiles, with benchmarking against the identified volcanic events, resulting in an estimated dating uncertainty of ±1 year. The measurement of the isotopic composition and the dating of the extension core B26-2012 was conducted at Copenhagen University. The annual mean $\delta^{18}O$ time series of the extension cores were calculated from the raw $\delta^{18}O$ data over depth and the depth–age relationship, as for the NEGIS core based on the published NEGIS raw data and depth–age relationship. Accumulation rates for the extension cores were derived from the density measurements and the depth–age relationship.

NGT record extensions

We extend the existing isotope records at the sites B18, B21, B23, B26 and NGRIP, which end in the mid-1990s, with the respective new records until the year 2011. To investigate the reliability of this approach we statistically analysed the overlap period between old and new records considering different running mean filtering window sizes from 1 to 21 years (Extended Data Table 3). The correlation of the annual mean data within the overlap period is somewhat low (≤0.25), probably owing to the strong relative

contribution of stratigraphic noise in single records[56], but the correlation systematically increases with increasing window size, with the best correlation observed for 11-year and 21-year filtered data, making the new records faithful representations for the old ones on these timescales.

To account for possible influences from different drilling or measurement techniques, we subtract from the new records the difference in mean isotopic composition within the overlap period (Extended Data Table 3). Starting from the earliest date of the overlap period onwards, the old records are then replaced by the new ones, extending the original records into the year 2011 (2010 for B26), resulting in an effective dataset of 16 $\delta^{18}O$ anomaly records.

The NGT-2012 isotope stack

We compile our effective dataset of 16 $\delta^{18}O$ anomaly records into a single stack by calculating the simple arithmetic average $\delta^{18}O$ value for each year ('NGT-2012' stack; Fig. 1). Owing to the different lengths of the firn cores and the different accumulation rates at the drill sites, the total number of firn cores included in the stack changes through time (Fig. 1a). To limit the influence of a very low number of records, we restrict our analyses to the time span 1000–2011, for which the NGT-2012 stack includes a minimum of four records (12 on average).

Temperature calibration of the NGT-2012 stack

For the conversion from isotopic composition to temperature, linear calibrations exist based either on the relationship of observed present-day spatial gradients in surface snow isotopic composition and temperature (spatial calibration) or on temporal gradients observed at a single site (temporal calibration). Because we work with anomaly time series, we only need to apply a calibration slope (°C per ‰). Here, we use the spatial slope for Greenland of 1/0.67 °C per ‰ (ref. [25]) and compare the results to those obtained from using the Holocene temporal slope of 2.1 °C per ‰ from a previous work[57] and the temporal slope for the NEEM site (estimated over 1979–2007) of 1/1.1 °C per mille[20], equivalent to a range of ±40% around the spatial slope. We do not apply any Last Glacial Maximum (LGM)–Holocene temporal slope, as it is not representative for present-day

conditions[58] owing to a different seasonality in precipitation or moisture source during the LGM[59,60].

For the spatial slope, we find the last 11 years of the NGT-2012 stack to be on average 1.7 ± 0.4 °C (±1 standard error) warmer than the 1961–1990 reference period and 1.5 ± 0.4 °C warmer than the twentieth century (1901–2000). These values correspond to temperature differences of 2.4 ± 0.6 °C and 2.1 ± 0.6 °C for the temporal slope from ref. [57] and to 1.0 ± 0.2 °C and 0.9 ± 0.2 °C for the NEEM temporal slope[20], showing that the overall uncertainty in the temperature difference when including the uncertainty on the calibration slope is significantly higher than the estimated standard error of the temperature difference itself.

Firn diffusion

Firn diffusion smooths the isotope signal with increasing strength as a function of time since deposition, described by the diffusion length, until the diffusion process ceases when the firn reaches the density of ice at bubble close-off. As a result, large amplitudes at the surface are damped with increasing depth. We model the diffusion length at each firn core site as a function of depth based on the standard theory for firn diffusion[61], using constant values for the local parameters of annual mean temperature, accumulation rate, surface pressure and surface snow density, as published in the literature[5,52,62,63,64,65,66,67]. To convert the diffusion lengths from depth into time units, we adopt the Herron–Langway densification model[68].

Owing to the increasing diffusion length, past events of elevated (warm) isotope values might have been stronger initially, that is, prior to diffusion. To assess the impact of firn diffusion on the distribution of the isotopic composition in the NGT-2012 stack, we artificially forward-diffuse each record as if it had been already completely densified to ice by applying a time-dependent differential diffusion length $\sigma(t)$ of

$$\sigma (t)=\sqrt{{\sigma }_{{\rm{ice}}}^{2}-{\sigma }_{{\rm{local}}}^{2}(t)},$$

in which σ_{ice} is the modelled diffusion length at the firn–ice transition and $\sigma_{local}(t)$ is the modelled diffusion length at each time point of the record.

Spectral analysis

We apply spectral analyses to the isotope records to derive timescale-dependent estimates (power spectral density) of the common climate signal and of the independent local noise, following a previous method[21]. The resulting signal and noise spectra are integrated to compute first the signal-to-noise variance ratio (SNR) as a function of the time resolution of the records and second, based on this, the corresponding expected correlation with the common signal as a function of the number of records averaged[21]. Because the spectral analysis relies on a fixed number of records for each time point, we restrict the analysis to the time span 1505–1978, which includes 14 of the 16 available records, and which is a trade-off between using many records and covering a sufficiently long time period for the spectral analysis. No diffusion correction is applied to the spectra, but we estimate the timescale range that is most affected by diffusion by determining the critical frequency at which the spectral diffusion transfer function takes a value of $1/e \approx 0.37$. This frequency depends on the value of the diffusion length; adopting the maximum of the estimated diffusion lengths across all isotope records and all observation points in time yields a critical frequency of ~1/7 $year^{-1}$ above which the spectra should be interpreted with care (Extended Data Fig. 1).

We find a distinct local maximum in the variability of the common signal (increased spectral power compared to a power-law background) around the 20-year period (Extended Data Fig. 1a), indicating enhanced climate variability at these timescales. The timescale-dependent estimate of the SNR increases continuously towards longer timescales and scales with the number of records averaged (Extended Data Fig. 1b), ranging from 3.4 at 11-year timescales for the average number of records in the NGT-2012 stack of $n = 12$, compared to 1.1 for $n = 4$ (minimum number) and 4.6 for $n = 16$ (maximum number), to 5.8 at the 100-year period (1.9–7.7). These values correspond to an expected correlation with the common signal at 11-year timescales of 0.73 for averaging $n = 4$ records and ≥0.85 for averaging $n \geq 12$ records (Extended Data Fig. 1b).

We estimate the magnitude-squared coherence between time series to assess their linear relationship as a function of timescale using the smoothed periodogram. Confidence levels are obtained by replacing the original time series with AR1 red-noise surrogate time series with the same autocorrelation and using the frequency averaged $p = 0.95$ sample quantile of $n = 1,000$ realizations.

Running mean filter and boundary constraints

Prior to the merging of the extended isotope records and the building of the NGT-2012 stack, we apply a running mean filter to each individual record using a window size of 11 years, which is based on the observed correlation within the overlap period of the extended isotope records (Extended Data Table 3), the reasonably high (~0.3) signal-to-noise ratio of a single record at the 11-year timescale (Extended Data Fig. 1b), and avoiding the range of timescales strongly affected by diffusion (Extended Data Fig. 1). To avoid data loss at the time series boundaries from applying the running mean filter, we adopt the 'minimum slope' boundary constraint[69], which is suited for the smoothing of potentially non-stationary time series and which is considered to modestly underestimate the behaviour of the time series near the boundaries in the presence of a long-term trend[69].

Pre-industrial distribution and comparison to the 2001–2011 time interval

To place the elevated isotope values of the recent 2001–2011 time interval into the historical context of our record, we compute the histogram of the 11-year running mean filtered values of the pre-industrial period (1000–1800). We fit a Gaussian distribution to the histogram, and compare this distribution to the block-averaged value of the recent time interval (Fig. 4a and Extended Data Fig. 3a), finding an extremely low probability for the recent value to occur under the pre-industrial distribution ($P = 1.82 \times 10^{-5}$, Extended Data Table 2).

The NGT-2012 accumulation rate stack

For an NGT-2012 accumulation rate stack (Extended Data Fig. 6a), we compiled accumulation rate records from the extension cores (B18-2012, B21-2012, B23-2012, B26-2012 and NGRIP-2012) as well as from the cores B16, B18, B21, B26, B29 and NEEM; the data of the remaining cores could not be used owing to insufficient quality. From a spectral analysis equivalent to the one applied to the isotopic data we find a timescale-dependent SNR for the accumulation rate data (Extended Data Fig. 6b) that is much lower (up to a factor of ~3) than the SNR of the isotopic data. One reason for such a low SNR is the strong spatial variability in local accumulation rates, which affects the accumulation rate reconstructions as local noise, but which can also create long-term artefacts if the spatial variability upstream of the ice-core site affects the down-core record by ice flow. As a result, for NGT-2012 we here use a simple stack of averaging across all available accumulation rate records without first merging the three available pairs of old and extension records, as is done with the isotope data, because the much higher noise level of the accumulation rate data rendered this approach inapplicable. The NGT-2012 isotope and accumulation rate stacks exhibit a low correlation of $R = 0.23$ ($P = 0.05$, $n = 512$) over 1500–2011, as can be expected from the low SNR of the accumulation rate data, without any statistically significant linear relationship (Extended Data Fig. 6a). Even though the NGT-2012 accumulation rate can be seen to have been increasing since 2000, similar to the isotopic data, this time interval is too short to derive any general relationship. In addition, the 2001–2011 block-averaged accumulation rate is not exceptional in the context of the pre-industrial values (Extended Data Fig. 6c), which could be due to noise in the reconstruction or a low sensitivity of the accumulation rate to the recent climate change.

Comparison with Arctic 2k data

We compare the NGT-2012 isotope stack with the Arctic 2k temperature reconstruction (1–2000)[13]. To cover the full time span of the NGT-2012 stack, we extend the published Arctic 2k record to 2011 with the HadCRUT near-surface instrumental temperature dataset version 5.0.1.0[70] by using the global gridded ensemble mean field of monthly anomalies, computing the annual mean anomalies for each grid cell, taking the area-weighted mean across all grid cells between 60° N and 90° N, and extending the annual

Arctic 2k dataset with these data from the year 2001 onwards (Fig. 1a and Extended Data Fig. 4b).

The overall correlation between the extended Arctic 2k reconstruction and the NGT-2012 stack after applying the 11-year running mean filter is $R = 0.65$ ($P < 0.001$, $n = 1{,}012$; $R = 0.58$, $P < 0.001$, $n = 2{,}001$ without extension); the correlation over 1901–2011 is $R = 0.66$ ($P < 0.01$, $n = 111$) but only 0.28 ($P = 0.17$, $n = 100$) without extension. A running correlation with a 101-year window size yields a mean correlation of 0.51 and shows variations that overall are within the range expected from surrogate data ($P = 0.84$ that the variations are to be expected by chance), but with unusually low correlation values for the twentieth century (Extended Data Fig. 4c).

The Arctic 2k reconstruction includes the original isotope records from GISP2, GRIP, NGRIP, B16, B18 and B21, which are also used in our compilation. To assess the extent to which these records contribute to the overall Arctic 2k temperature reconstruction, we correlate our extended versions for each of these records with the Arctic 2k record, yielding correlations in the range from 0 to 0.5 (specifically, GRIP: 0.00, GISP2: 0.29, NGRIP: 0.19, B16: 0.39, B18: 0.37 and B21: 0.49; $n = 1{,}001$). The record from location B21 shows the highest correlation, which is the farthest north and at the lowermost elevation. However, the overall low correlation of these records indicates that their contribution to the Arctic 2k record itself is limited.

Comparison with instrumental temperature data

We correlate the NGT-2012 isotope stack with nearby instrumental temperature data from the weather stations Upernavik, Pituffik and Danmarkshavn from the Danish Meteorological Institute[16] covering the time period 1873–2011, applying the same 11-year running mean filter to the instrumental temperature data as to the isotope record (Extended Data Fig. 2). We obtain correlation coefficients of $R = 0.87$ (Pituffik, 1948–2011), $R = 0.75$ (Upernavik, 1901–2011) and $R = 0.85$ (Danmarkshavn, 1949–2011) (all $P < 0.005$), which are in the range of expected correlations from our spectral analysis, supporting the interpretation of the isotope stack as a temperature signal for the area. We note that including the instrumental data from

Upernavik prior to 1901 yields a weaker correlation with the NGT-2012 stack, which could be due to limitations of the instrumental data or a weaker representativity of the instrumental record for the area of our firn cores.

Comparison with reanalysis data

We compute the point correlations of the near-surface temperature field from the Twentieth Century Reanalysis version 3 (20CRv3)[23,24] dataset in the time window 1836–2000 for all grid cells ≥50° N with the NGT-2012 $\delta^{18}O$ anomalies and with the Arctic 2k reconstructed temperature anomalies, using both 11-year running mean as well as annual mean data (Fig. 2 and Extended Data Fig. 5). We specifically rely here on reanalysis data, because no direct instrumental temperature observations exist on the Greenland Ice Sheet and thus observational datasets, such as HadCRUT[71], practically interpolate sea-level-based coastal station data over the ice sheet, leading to spurious correlations. The analyses show that the NGT-2012 record is strongly correlated with the reanalysis temperature over the Greenland Ice Sheet but that the Arctic 2k reconstruction only exhibits nonsignificant correlations there. Although here we focus our analyses on 11-year running mean anomalies, this result is largely robust also for annual mean values.

MAR3.5.2 surface mass balance and temperature estimates

Greenland meltwater run-off is obtained as a component of the surface mass balance (SMB) output of the regional climate model MAR3.5.2 (Modèle Atmosphérique Régional; version 3.5.2)[22]. Meltwater run-off refers to meltwater production minus meltwater refreezing, deposition and retention. The MAR3.5.2 simulation used here is forced in six-hourly intervals at its lateral boundaries with Twentieth Century Reanalysis version 2 (20CRv2)[23] for the period 1871–2012, and provides 20-km horizontal resolution. This model output is part of a larger number of twentieth-century reconstructions of the Greenland Ice Sheet SMB with MAR3.5.2, forced by various different atmospheric reanalysis datasets[22]. The 20CRv2 forcing is the ensemble mean of a 56-member experimental reanalysis with spatial resolution of 2.0°, assimilating only surface pressure, monthly sea surface temperature and sea ice cover[22].

For the period 1980–2010, MAR3.5.2 forced by 20CRv2 has been shown to exhibit a warm temperature bias (~1 °C) compared to simulations driven by ECMWF Interim reanalysis[72]. However, for the annual meltwater run-off anomalies with respect to 1961–1990 considered in this study, we find that MAR3.5.2/20CRv2 is in good agreement with the latest version MAR3.12 forced by the latest reanalysis (for example, ERA5[73]; Extended Data Fig. 7), within the common period 1950–2012. Even though it is not possible to directly measure mass changes due to meltwater run-off with satellites, we estimate the meltwater run-off anomaly by subtracting net accumulation (snowfall minus sublimation and evaporation) obtained from MAR3.12/ERA5 and ice dynamic discharge obtained from InSAR[42,74] from the GRACE/GRACE-FO annual mass balance with breakpoint January of each year. The results show that the annual variation of the mass budget based on MAR3.12/ERA-5 is consistent with GRACE/GRACE-FO, as is the budget when replacing the meltwater run-off from MAR3.12/ERA5 with the MAR3.5.2/20CRv2 estimates (Extended Data Fig. 7).

For our study, we base the Greenland meltwater run-off anomalies and 2-m surface air temperature data on monthly estimates from MAR3.5.2. The monthly temperature data are sampled at the grid cells closest to the NGT-2012 ice-core locations, averaged across these cells and then averaged to annual mean values; the meltwater run-off data are integrated over the contiguous ice sheet and then cumulated to annual values. Anomalies are calculated with respect to the reference period 1961–1990, which is, first, the commonly used reference period in mass balance studies of the Greenland Ice Sheet[75], and second, synchronous to the one used for the NGT-2012 and Arctic 2k time series. Finally, the same 11-year running mean filter is applied to the annual temperature values as to the NGT-2012 isotope record, yielding a correlation with the filtered NGT-2012 record over the common time period 1871–2011 of $R = 0.76$ ($P < 0.01$, $n = 141$). Likewise, the correlation of the filtered MAR3.5.2 meltwater run-off anomaly with NGT-2012 is $R = 0.62$ ($P < 0.01$, $n = 141$).

Comparison with Greenland Blocking Index

We compare the Greenland Blocking Index (GBI)[29] time series to the NGT-2012 temperature and MAR3.5.2 meltwater run-off data over their common

time periods. Using 11-year running mean filtered data, the correlation between NGT-2012 and annual GBI is $R = 0.63$ ($P < 0.005$, $n = 161$) and between meltwater run-off and annual GBI it is $R = 0.80$ ($P < 0.001$, $n = 141$). Replacing the annual GBI data with the average GBI for summer (months June, July, August), the correlation with meltwater run-off is $R = 0.91$ ($P < 0.001$, $n = 141$). The correlations are robust also for the unfiltered annual mean values, with correlations of $R = 0.39$ ($P < 0.001$, $n = 161$), $R = 0.56$ ($P < 0.001$, $n = 141$), and $R = 0.67$ ($P < 0.001$, $n = 141$), respectively.

Significance of correlation between filtered time series

Significance values for the correlation estimates between two running-mean filtered time series (hereafter, 'data' and 'signal') are derived from a Monte Carlo sampling approach, in which $n = 10{,}000$ realizations ($n = 1{,}000$ for the correlation maps) of random surrogate data are created with the same AR1 autocorrelation structure as the original (that is, unfiltered) data, filtered with the same running mean filter as the original data, and correlated with the filtered signal. The significance of the observed correlation between filtered data and signal is then obtained from the fraction of surrogate correlations that exceed the observed correlation.

The significance of the running correlation between filtered data and signal is estimated following a method previously described[76]. The correlation between the unfiltered data and signal is used to create $n = 10{,}000$ random surrogate time series, which exhibit on average the same correlation with the signal as the original data. Surrogate data and the signal are filtered and the running correlation between them is computed. From these surrogate running correlations, we report the local 2.5–97.5% quantiles, and, by expressing the correlation values in terms of z values[76], the overall significance of the variations in the observed running correlation is obtained from the fraction of maximum z value differences for the surrogate data which exceed the maximum z value difference of the observation.

Sensitivity of probability results

To test the robustness of the found probability for the recent isotope value to occur under the pre-industrial distribution we investigate different variants

of creating and analysing the NGT-2012 stack. Specifically, we compare our results based on the main NGT-2012 stack (Fig. 4a) to those obtained for building (1) the NGT-2012 stack from artificially fully forward-diffused data, (2) a stack with a fixed number ($n = 5$) of records through time, (3) a stack from simply averaging across all available isotope records without merging old and new records, (4) as before but including full artificial forward diffusion, and (5) the NGT-2012 stack without adjusting for the difference in mean value within the overlap interval of old and new records (Extended Data Fig. 3 and Extended Data Table 2). All these variants lead to similar probability values for the recent value in the range of $P = 1.8–2.6 \times 10^{-5}$ (Extended Data Table 2). For the main NGT-2012 stack, we additionally vary the length of the running mean filter window and the length of the pre-industrial period (shifting it to maximum 1900), which does not affect the probability value notably (all $P \leq 10^{-5}$), except for a running mean filter window of 7, 9 and 21 years ($P \approx 10^{-4}$; Extended Data Table 2). Finally, we adjust the range of the recent period by shifting it into the past in steps of 1 year. This systematically increases the probability value by nearly two orders of magnitude (Extended Data Table 2), which is expected because the earlier ranges correspond to significantly less elevated isotope values in the NGT-2012 time series (Fig. 1a). We note that the marginal effect of firn diffusion is due to the relatively high accumulation rates at the sites[5] ($\gtrsim$100 kg m^{-2} year^{-1}), resulting in small differential diffusion lengths (≤1 year in time units), which have a strong impact on annual and interannual isotope values but only a negligible effect on longer timescales.

Reconstruction of pre-industrial meltwater run-off distribution

We reconstruct the distribution of meltwater run-off anomalies for the time period of NGT-2012 based on the linear relationship between the NGT-2012 temperatures $T_{\rm core}$ and MAR3.5.2 meltwater run-off $M_{\rm MAR}$ anomalies for the period 1871 and 2012,

$${M}_{{\rm{MAR}}}^{1871\mbox{--}2012}={T}_{{\rm{core}}}^{1871\mbox{--}2012}\beta +{\epsilon },$$

where β is the linear regression coefficient and ϵ represents uncertainties. We estimate $\hat{\beta }$ and its variance ${\rm{var}}(\hat{\beta })$ using least-squares adjustment, with the assumption of uniform uncertainties in ${M}_{{\rm{MAR}}}^{1871\mbox{--}2012}$. The reconstructed meltwater run-off $\hat{M}$ for the pre-industrial time period (PI; 1000–1800) based on $T_{\rm core}$ is then obtained as

$$\hat{M}={T}_{{\rm{core}}}^{{\rm{PI}}}\,\hat{\beta }.$$

To account for uncertainties related to the parameter estimate, as well as the post-fit residual, we calculate the variance of the melt run-off reconstruction as

$${\rm{var}}(\hat{M})={\rm{var}}({M}_{{\rm{MAR}}}^{1871\mbox{--}2012}-{T}_{{\rm{core}}}^{1871\mbox{--}2012}\hat{\beta })+{\rm{var}}(\hat{\beta }){T}_{{\rm{core}}}^{{\rm{PI}}},$$

using a Monte Carlo approach involving 10,000 random samples.

To derive the two-dimensional distribution of pre-industrial meltwater run-off versus temperature data, we create a 2D grid with 50 bins in each direction spanning the range $[T_1, T_2]$ and $\hat{\beta }[{T}_{1},{T}_{2}]$, where $T_1 = -4$ °C and $T_2 = 4$ °C, and count the number of realizations that fall into each of the bins. The meltwater reconstruction based on the full time period covered by NGT-2012 is obtained by ${\hat{M}}_{{\rm{full}}}={T}_{{\rm{core}}}\,\hat{\beta }$.

We note that the finding of the 2001–2011 decade being outside of the pre-industrial distribution is partly a result of this linear reconstruction from the NGT-2012 data, where the 2001–2011 decade is exceptional. The overall run-off is physically not directly linked to temperature, but (1) here we find a linear relationship over the 1871–2012 time period between NGT-2012 and Greenland meltwater run-off and (2) we know that the area affected by melt is changing with changing temperature (increasing under warming conditions). Therefore, we assume that the overall response of the meltwater run-off to changing temperature is linear and thus a linear reconstruction is feasible.

Data availability

All ice-core stable isotope data used in this study are publicly available in online data repositories from the links given in Extended Data Table 1. The accumulation rate data of the extension cores are available from the respective links for the isotope data; the accumulation rate data of the other used cores can be obtained from the exNGT software project archived under https://doi.org/10.5281/zenodo.7178657. The reconstructed meltwater run-off time series is publicly available at https://doi.org/10.7910/DVN/XQMAY6. Instrumental temperature data for Greenland are available from the Danish Meteorological Institute and were obtained from Polar Portal at http://polarportal.dk/en/news/news/historical-weather-and-climatedata-for-greenland/. Twentieth Century Reanalysis version 3 data are provided by the NOAA/OAR/ESRL PSL, Boulder, CO, USA, at https://www.psl.noaa.gov/data/gridded/data.20thC_ReanV3.html. HadCRUT5 data were obtained from the Met Office Hadley Centre at https://www.metoffice.gov.uk/hadobs/hadcrut5/index.html. For MARv3.5.2 output we refer to the original publication providing the download link. Source data are provided with this paper.

Code availability

Software to reproduce the analyses is available as R code hosted in the public Git repository exNGT at https://github.com/EarthSystemDiagnostics/exNGT, a snapshot of this code (exNGT version 1.0.0) at the time of publication is archived under https://doi.org/10.5281/zenodo.7178657.

References

1. Pattyn, F. et al. The Greenland and Antarctic ice sheets under 1.5 °C global warming. *Nat. Clim. Change* **8**, 1053–1061 (2018).

2. Hanna, E., Mernild, S. H., Cappelen, J. & Steffen, K. Recent warming in Greenland in a long-term instrumental (1881–2012) climatic context:

I. Evaluation of surface air temperature records. *Environ. Res. Lett.* **7**, 045404 (2012).

3. Orsi, A. J. et al. The recent warming trend in North Greenland. *Geophys. Res. Lett.* **44**, 6235–6243 (2017).

4. Kobashi, T. et al. High variability of Greenland surface temperature over the past 4000 years estimated from trapped air in an ice core. *Geophys. Res. Lett.* https://doi.org/10.1029/2011GL049444 (2011).

5. Weißbach, S. et al. Spatial and temporal oxygen isotope variability in northern Greenland – implications for a new climate record over the past millennium. *Clim. Past* **12**, 171–188 (2016).

6. Allen, M. R. et al. in *Global Warming of 1.5 °C. An IPCC Special Report on the Impacts of Global Warming of 1.5 °C Above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty* (eds Masson-Delmotte, V. et al.) Ch. 1 (Cambridge Univ. Press, 2018).

7. Serreze, M. C. & Barry, R. G. Processes and impacts of Arctic amplification: a research synthesis. *Global Planet. Change* **77**, 85–96 (2011).

8. Maturilli, M. & Kayser, M. Arctic warming, moisture increase and circulation changes observed in the Ny-Ålesund homogenized radiosonde record. *Theor. Appl. Climatol.* **130**, 1–17 (2017).

9. Marotzke, J. & Forster, P. M. Forcing, feedback and internal variability in global temperature trends. *Nature* **517**, 565–570 (2015).

10. Laepple, T. & Huybers, P. Ocean surface temperature variability: large model–data differences at decadal and longer periods. *Proc. Natl Acad. Sci.* **111**, 16682–16687 (2014).

11. Swanson, K. L., Sugihara, G. & Tsonis, A. A. Long-term natural variability and 20th century climate change. *Proc. Natl Acad. Sci.* **106**,

16120–16123 (2009).

12. Christiansen, B. & Ljungqvist, F. C. Challenges and perspectives for large-scale temperature reconstructions of the past two millennia. *Rev. Geophys.* **55**, 40–96 (2017).

13. McKay, N. P. & Kaufman, D. S. An extended Arctic proxy temperature database for the past 2,000 years. *Sci. Data* **1**, 140026 (2014).

14. Abram, N. J. et al. Early onset of industrial-era warming across the oceans and continents. *Nature* **536**, 411–418 (2016).

15. McGrath, D., Colgan, W., Bayou, N., Muto, A. & Steffen, K. Recent warming at Summit, Greenland: global context and implications. *Geophys. Res. Lett.* **40**, 2091–2096 (2013).

16. Cappelen, J. (ed.) *Greenland – DMI Historical Climate Data Collection 1784–2013*. Technical Report 14-04 (Danish Meteorological Institute, 2014).

17. Chylek, P., Box, J. E. & Lesins, G. Global warming and the Greenland Ice Sheet. *Clim. Change* **63**, 201–221 (2004).

18. Box, J. E. & Cohen, A. E. Upper-air temperatures around Greenland: 1964–2005. *Geophys. Res. Lett.* https://doi.org/10.1029/2006GL025723 (2006).

19. MacFerrin, M. et al. Rapid expansion of Greenland's low-permeability ice slabs. *Nature* **573**, 403–407 (2019).

20. Masson-Delmotte, V. et al. Recent changes in north-west Greenland climate documented by NEEM shallow ice core data and simulations, and implications for past-temperature reconstructions. *Cryosphere* **9**, 1481–1504 (2015).

21. Münch, T. & Laepple, T. What climate signal is contained in decadal to centennial scale isotope variations from Antarctic ice cores? *Clim. Past* **14**, 2053–2070 (2018).

22. Fettweis, X. et al. Reconstructions of the 1900–2015 Greenland ice sheet surface mass balance using the regional climate MAR model. *Cryosphere* **11**, 1015–1033 (2017).

23. Compo, G. P. et al. The Twentieth Century Reanalysis Project. *Q. J. R. Meteorol. Soc.* **137**, 1–28 (2011).

24. Slivinski, L. C. et al. Towards a more reliable historical reanalysis: improvements for version 3 of the Twentieth Century Reanalysis system. *Q. J. R. Meteorol. Soc.* **145**, 2876–2908 (2019).

25. Dansgaard, W. Stable isotopes in precipitation. *Tellus* **16**, 436–468 (1964).

26. Fyfe, J. C. et al. One hundred years of Arctic surface temperature variation due to anthropogenic influence. *Sci. Rep.* **3**, 2645 (2013).

27. Cassou, C. et al. Decadal climate variability and predictability: challenges and opportunities. *Bull. Am. Meteorol. Soc.* **99**, 479–490 (2018).

28. Polashenski, C. et al. Observations of pronounced Greenland ice sheet firn warming and implications for runoff production. *Geophys. Res. Lett.* **41**, 4238–4246 (2014).

29. Hanna, E., Cropper, T. E., Hall, R. J. & Cappelen, J. Greenland Blocking Index 1851–2015: a regional climate change signal. *Int. J. Climatol.* **36**, 4847–4861 (2016).

30. Overland, J. et al. The melting Arctic and midlatitude weather patterns: are they connected? *J. Clim.* **28**, 7917–7932 (2015).

31. Johannessen, O. M., Kuzmina, S. I., Bobylev, L. P. & Miles, M. W. Surface air temperature variability and trends in the Arctic: new amplification assessment and regionalisation. *Tellus A* **68**, 28234 (2016).

32. Venegas, S. A. & Mysak, L. A. Is there a dominant timescale of natural climate variability in the Arctic? *J. Clim.* **13**, 3412–3434 (2000).

33. Kobashi, T., Goto-Azuma, K., Box, J. E., Gao, C. C. & Nakaegawa, T. Causes of Greenland temperature variability over the past 4000 yr: implications for northern hemispheric temperature changes. *Clim. Past* **9**, 2299–2317 (2013).

34. Rimbu, N. & Lohmann, G. Winter and summer blocking variability in the North Atlantic region – evidence from long-term observational and proxy data from southwestern Greenland. *Clim. Past* **7**, 543–555 (2011).

35. Rimbu, N., Lohmann, G., Werner, M. & Ionita, M. Links between central Greenland stable isotopes, blocking and extreme climate variability over Europe at decadal to multidecadal time scales. *Clim. Dyn.* **49**, 649–663 (2017).

36. Hahn, L., Ummenhofer, C. C. & Kwon, Y.-O. North Atlantic natural variability modulates emergence of widespread Greenland melt in a warming climate. *Geophys. Res. Lett.* **45**, 9171–9178 (2018).

37. McLeod, J. T. & Mote, T. L. Linking interannual variability in extreme Greenland blocking episodes to the recent increase in summer melting across the Greenland ice sheet. *Int. J. Climatol.* **36**, 1484–1499 (2016).

38. Flowers, G. E. Hydrology and the future of the Greenland Ice Sheet. *Nat. Commun.* **9**, 2729 (2018).

39. van den Broeke, M. R. et al. On the recent contribution of the Greenland ice sheet to sea level change. *Cryosphere* **10**, 1933–1946 (2016).

40. Nienow, P. W., Sole, A. J., Slater, D. A. & Cowton, T. R. Recent advances in our understanding of the role of meltwater in the Greenland ice sheet system. *Curr. Clim. Change Rep.* **3**, 330–344 (2017).

41. WCRP Global Sea Level Budget Group. Global sea-level budget 1993–present. *Earth Syst. Sci. Data* **10**, 1551–1590 (2018).

42. Sasgen, I. et al. Return to rapid ice loss in Greenland and record loss in 2019 detected by the GRACE-FO satellites. *Commun. Earth Environ.*

1, 8 (2020).

43. Ohmura, A. Physical basis for the temperature-based melt-index method. *J. Appl. Meteorol.* **40**, 753–761 (2001).

44. Machguth, H. et al. Greenland meltwater storage in firn limited by near-surface ice formation. *Nat. Clim. Change* **6**, 390–393 (2016).

45. Mernild, S. H. & Liston, G. E. Surface melt extent for the Greenland Ice Sheet, 2011. *Geogr. Tidsskr.* **112**, 84–88 (2012).

46. Graeter, K. A. et al. Ice core records of west Greenland melt and climate forcing. *Geophys. Res. Lett.* **45**, 3164–3172 (2018).

47. Lenaerts, J. T. M. et al. Representing Greenland ice sheet freshwater fluxes in climate models. *Geophys. Res. Lett.* **42**, 6373–6381 (2015).

48. Fettweis, X. et al. Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR. *Cryosphere* **7**, 469–489 (2013).

49. Bamber, J., van den Broeke, M., Ettema, J., Lenaerts, J. & Rignot, E. Recent large increases in freshwater fluxes from Greenland into the North Atlantic. *Geophys. Res. Lett.* https://doi.org/10.1029/2012GL052552 (2012).

50. Bakker, P. et al. Fate of the Atlantic Meridional Overturning Circulation: strong decline under continued warming and Greenland melting. *Geophys. Res. Lett.* **43**, 12252–12260 (2016).

51. Buchardt, S. L., Clausen, H. B., Vinther, B. M. & Dahl-Jensen, D. Investigating the past and recent $\delta^{18}O$–accumulation relationship seen in Greenland ice cores. *Clim. Past* **8**, 2053–2059 (2012).

52. Vinther, B. M. et al. A synchronized dating of three Greenland ice cores throughout the Holocene. *J. Geophys. Res. Atmos.* https://doi.org/10.1029/2005JD006921 (2006).

53. Rasmussen, S. O. et al. A first chronology for the North Greenland Eemian Ice Drilling (NEEM) ice core. *Clim. Past* **9**, 2713–2730 (2013).

54. Freitag, J., Wilhelms, F. & Kipfstuhl, S. Microstructure-dependent densification of polar firn derived from X-ray microtomography. *J. Glaciol.* **50**, 243–250 (2004).

55. van Geldern, R. & Barth, J. A. C. Optimization of instrument setup and post-run corrections for oxygen and hydrogen stable isotope measurements of water by isotope ratio infrared spectroscopy (IRIS). *Limnol. Oceanogr. Meth.* **10**, 1024–1036 (2012).

56. Münch, T., Kipfstuhl, S., Freitag, J., Meyer, H. & Laepple, T. Regional climate signal vs. local noise: a two-dimensional view of water isotopes in Antarctic firn at Kohnen Station, Dronning Maud Land. *Clim. Past* **12**, 1565–1581 (2016).

57. Vinther, B. M. et al. Holocene thinning of the Greenland ice sheet. *Nature* **461**, 385–388 (2009).

58. Krinner, G. & Werner, M. Impact of precipitation seasonality changes on isotopic signals in polar ice cores: a multi-model analysis. *Earth Planet. Sci. Lett.* **216**, 525–538 (2003).

59. Jouzel, J. et al. Validity of the temperature reconstruction from water isotopes in ice cores. *J. Geophys. Res. Oceans* **102**, 26471–26487 (1997).

60. Jouzel, J., Hoffmann, G., Koster, R. D. & Masson, V. Water isotopes in precipitation: data/model comparison for present-day and past climates. *Quat. Sci. Rev.* **19**, 363–379 (2000).

61. Johnsen, S. J. C. et al. in *Physics of Ice Core Records* (ed. Hondoh, T.) 121–140 (Hokkaido Univ. Press, 2000).

62. Vallelonga, P. et al. Initial results from geophysical surveys and shallow coring of the Northeast Greenland Ice Stream (NEGIS). *Cryosphere* **8**, 1275–1287 (2014).

63. Cuffey, K. M. & Clow, G. D. Temperature, accumulation, and ice sheet elevation in central Greenland through the last deglacial transition. *J. Geophys. Res. Oceans* **102**, 26383–26396 (1997).

64. Grootes, P. M., Stuiver, M., White, J. W. C., Johnsen, S. & Jouzel, J. Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature* **366**, 552–554 (1993).

65. Dahl-Jensen, D. et al. Eemian interglacial reconstructed from a Greenland folded ice core. *Nature* **493**, 489–494 (2013).

66. Schaller, C. F. et al. A representative density profile of the North Greenland snowpack. *Cryosphere* **10**, 1991–2002 (2016).

67. Zuhr, A. M., Münch, T., Steen-Larsen, H. C., Hörhold, M. & Laepple, T. Local-scale deposition of surface snow on the Greenland ice sheet. *Cryosphere* **15**, 4873–4900 (2021).

68. Herron, M. M. & Langway, C. C. Firn densification: an empirical model. *J. Glaciol.* **25**, 373–385 (1980).

69. Mann, M. E. On smoothing potentially non-stationary climate time series. *Geophys. Res. Lett.* https://doi.org/10.1029/2004GL019569 (2004).

70. Morice, C. P. et al. An updated assessment of near-surface temperature change from 1850: the HadCRUT5 data set. *J. Geophys. Res. Atmos.* **126**, e2019JD032361 (2021).

71. Cowtan, K. & Way, R. G. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Q. J. R. Meteorol. Soc.* **140**, 1935–1944 (2014).

72. Dee, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).

73. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

74. King, M. D. et al. Seasonal to decadal variability in ice discharge from the Greenland Ice Sheet. *Cryosphere* **12**, 3813–3825 (2018).

75. van den Broeke, M. et al. Partitioning recent Greenland mass loss. *Science* **326**, 984–986 (2009).

76. van Oldenborgh, G. J. & Burgers, G. Searching for decadal variations in ENSO precipitation teleconnections. *Geophys. Res. Lett.* https://doi.org/10.1029/2005GL023110 (2005).

Acknowledgements

The re-drilling of the NGT ice cores in 2011 and 2012 was supported by the NEEM ice-core project, which is directed and organized by the Center of Ice and Climate at the Niels Bohr Institute and US NSF, Office of Polar Programs. It is supported by funding agencies and institutions in Belgium (FNRS-CFB and FWO), Canada (NRCan/GSC), China (CAS), Denmark (FIST), France (IPEV, CNRS/INSU, CEA and ANR), Germany (AWI), Iceland (RannIs), Japan (NIPR), South Korea (KOPRI), the Netherlands (NWO/ALW), Sweden (VR), Switzerland (SNF), the UK (NERC) and the USA (US NSF, Office of Polar Programs). We thank the Polar 5 team of the NGT 2012 campaign D. Steinhage, J. Binder, E. Bengtsson, J. Bayes and G. Cirtwill. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 716092) and was supported by Helmholtz funding through the Polar Regions and Coasts in the Changing Earth System (PACES) programme of the Alfred Wegener Institute as well as through the Helmholtz Research Field Earth and Environment Research Program Changing Earth – Sustaining our Future. We acknowledge support by the Open Access Publication Funds of Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung. Support for the Twentieth Century Reanalysis Project version 3 dataset is provided by the US Department of Energy, Office of Science Biological and Environmental Research, by the National Oceanic and Atmospheric Administration (NOAA) Climate Program Office, and by the NOAA Physical Sciences Laboratory. The main plots and numerical analyses in this study were carried out using the open-source software R: A Language and Environment for

Statistical Computing. We thank H. Oerter, who supervised the $\delta^{18}O$ measurements of the NGT ice cores. We thank X. Fettweis for publicly providing output of the regional climate model MAR.

Funding

Open access funding provided by Alfred-Wegener-Institut.

Author information

Authors and Affiliations

1. Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany

   M. Hörhold, S. Weißbach, S. Kipfstuhl, J. Freitag, I. Sasgen & G. Lohmann

2. Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Potsdam, Germany

   T. Münch & T. Laepple

3. Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

   B. Vinther

4. MARUM – Center for Marine Environmental Sciences and Faculty of Geosciences, University of Bremen, Bremen, Germany

   T. Laepple

Contributions

M.H., T.L. and T.M. designed the study and analyses and wrote the initial manuscript draft. T.L. and M.H. oversaw the research. M.H. and T.M. drafted the merging and stacking procedure of the re-drilled ice cores, T.M.

developed and wrote the R code. S.W. and J.F. conducted the stable water isotope measurements, J.F. provided the density data and accomplished the dating of the re-drilled ice cores, together with S.W. and T.M. S.K. retrieved the NGT and NGT-2012 cores, B.V. provided all data and dating for B26-2011. I.S. conducted all analysis with MAR and G.L. contributed to the analysis with climate data. All authors contributed to the manuscript.

Corresponding author

Correspondence to M. Hörhold.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Spectral analysis of the NGT-2012 isotope stack.

a, The power spectral density (PSD) of the spatially coherent signal common to all individual firn core records shows enhanced variability around the 20-year period relative to a power-law-type increase towards longer timescales. **b**, Estimated signal-to-noise ratio of the isotope data as a function of the timescale (left axis) and the corresponding correlation with the common signal (right axis). The thick black line shows the signal-to-noise ratio and the correlation for a stack of $N = 12$ records (the average number of records in the NGT-2012 stack) with the grey-shaded area indicating the range in values for record numbers from $N = 4$ (minimum number in the NGT-2012 stack) to $N = 16$ (maximum number). The red-shaded areas in **a** and **b** indicate the range of timescales strongly influenced by diffusion.

Extended Data Fig. 2 Comparison of NGT-2012 with instrumental and model data.

a, NGT-2012 $\delta^{18}O$ 11-yr running mean anomalies (black) for the period from 1901 to 2011 together with the 11-yr running mean air temperature anomalies from the three Greenlandic coastal weather stations[16] Pituffik (green), Upernavik (orange) and Danmarkshavn (purple) of the Danish Meteorological Institute (DMI). Correlation coefficients between NGT-2012 and the individual temperature records are $R = 0.87$ (Pituffik; $P < 0.005$, $n = 64$), $R = 0.75$ (Upernavik; $P < 0.005$, $n = 111$) and $R = 0.85$ (Danmarkshavn; $P < 0.005$, $n = 63$). **b**, Comparison of NGT-2012 11-yr running mean temperature anomalies (black), where grey shading denotes a ~40 % uncertainty of the temperature reconstruction obtained from the range of plausible calibration slopes (Methods), with the temperature anomalies from the regional climate model MAR3.5.2[22] averaged across the NGT-2012 sites (Methods) for the time period 1871–2011. **c**, As **b**, but for the comparison with MAR3.5.2 Greenland meltwater run-off anomalies. **d**, as **b** but for the comparison with the Greenland Blocking Index (GBI)[29]. Correlation coefficients between NGT-2012 temperature and MAR3.5.2 temperature, MAR3.5.2 meltwater run-off and GBI are $R = 0.76$ ($P < 0.001$, $n = 141$), $R = 0.62$ ($P < 0.01$, $n = 141$) and $R = 0.63$ ($P < 0.005$, $n = 161$), respectively. **e**, Scatter plot of NGT-2012 temperature anomalies versus GBI. **f**, Scatter plot of MAR3.5.2 meltwater run-off anomalies versus GBI (correlation $R = 0.80$, $P < 0.001$, $n = 141$). **g**, Scatter plot of MAR3.5.2 meltwater run-off

anomalies versus NGT-2012 temperature anomalies with the blue line indicating a least squares linear regression.

Extended Data Fig. 3 NGT-2012 pre-industrial probability density distribution for different analysis variants.

a–**d**, Compared are our main analysis for the NGT-2012 stack (**a** and Fig. 3) with different analysis variants: the NGT-2012 stack including full forward diffusion of the data (**b**; Methods); a stack built from a fixed number (N = 5) of records through time (**c**); and a simple stack built from averaging across all available isotope records without merging old and new records (**d**). **e**, As in **d** but including full forward diffusion. These variants of deriving and analysing the NGT-2012 stack have an only marginal influence on our results (Extended Data Table 3).

Extended Data Fig. 4 Comparison of NGT-2012 with Arctic 2k.

a, NGT-2012 $\delta^{18}O$ 11-yr running mean anomalies for the period from 1000 to 2011 with **b**, 11-yr running mean temperature anomalies relative to the mean in the 1961–1990 reference period from the Arctic 2k temperature reconstruction[13] extended by HadCRUT5[70] instrumental temperature data to the year 2011. **c**, Running correlation over consecutive 101-yr windows between the extended Arctic 2k temperature reconstruction[13] and the NGT-2012 $\delta^{18}O$ data over the time period 1000–2011. The dashed thin black lines show the 2.5–97.5% quantile range of running correlation results for n = 10,000 realizations of surrogate data with the same mean correlation to the Arctic 2k data as the NGT-2012 data, indicating that the observed correlation variations between NGT-2012 and Arctic 2k data are likely to be expected by chance (p = 0.84).

Extended Data Fig. 5 NGT-2012 and Arctic 2k point correlation with the Twentieth Century Reanalysis v3 (20CRv3) near-surface temperature field.

a, Point correlation between the 20CRv3 reanalysis field[23,24] of annual mean near-surface temperatures and the NGT-2012 annual mean $\delta^{18}O$ temperature

reconstruction time series. **b**, As in **a** but for the point correlation with the Arctic 2k annual mean temperature reconstruction time series. **c**, As in **a** but for the point correlation with the 20CRv3 reanalysis temperature area mean of the Greenland region. **d**, As in **a** but for the point correlation with the 20CRv3 reanalysis temperature area mean of the Arctic region (60–90° N). Correlations are calculated for the time period 1836–2000 (**a**,**b**) or 1836–2015 (**c**,**d**) for all reanalysis grid cells ≥50° N. Grid cells filled grey mark areas with nonsignificant correlation values ($P > 0.05$; $n = 165$ in **a**,**b**; $n = 180$ in **c**,**d**). All geographic map data are obtained from the 'rnaturalearth' package for the software R.

Extended Data Fig. 6 NGT-2012 accumulation rate data.

a, NGT-2012 accumulation rate and NGT-2012 $\delta^{18}O$ 11-yr running mean anomaly time series from this study for the time period 1500 to 2011. **b**, Estimated signal-to-noise ratio of the NGT-2012 accumulation rate data as a function of the timescale, where the thick black line shows the signal-to-noise ratio for a stack of $N = 6$ records (the average number of records in the NGT-2012 accumulation rate stack) with the grey-shaded area indicating the range in values for record numbers from $N = 4$ (minimum number in the accumulation rate stack) to $N = 11$ (maximum number). **c**, Histogram of the 11-yr running mean NGT-2012 accumulation rate values for the pre-industrial time period (1500–1800) together with a Gaussian fit (thick black line). Vertical black lines show the quantiles corresponding to probabilities of $p = 0.95$ (dash-dotted) and $p = 0.99$ (dashed), respectively. The 2001–2011 block-averaged NGT-2012 accumulation rate value is shown as a thick orange line.

Source data

Extended Data Fig. 7 Greenland meltwater run-off estimated from model and satellite data.

Shown are the meltwater run-off anomalies for MAR3.5.2[22] forced by 20CRv2[23] (1950–2012; blue), MAR3.12 forced by ERA5[73] (1950–2021; green) and as estimated based on GRACE/GRACE-FO[42,74] data (2002–2020 ad; black). Anomalies for MAR3.5.2 and MAR3.12 are with respect to

the years 1961–1990. GRACE/GRACE-FO meltwater run-off anomalies represent the annual mass change measured by GRACE/GRACE-FO minus the net snow accumulation from MAR3.12 and the ice-dynamic discharge from InSAR[74]. An offset was removed to reconcile the mean meltwater run-off anomaly from MAR3.12 and GRACE/GRACE-FO in the period 2002–2020. GRACE/GRACE-FO uncertainties (2σ) are propagated empirical uncertainties of the mass change only, and do not contain uncertainties associated with the subtracted components net accumulation and ice-dynamic discharge. The reconciliation of the mean mass budget from GRACE/GRACE-FO and details on the GRACE/GRACE-FO data as well as the surface mass balance models are provided in ref. [42].

Extended Data Table 1 Overview of firn core records included in the NGT-2012 stack
Extended Data Table 2 Probability analysis
Extended Data Table 3 Overlap statistics

Supplementary information

Peer review file

Source data

Source Data Fig. 1

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 4

Source Data Extended Data Fig. 6

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Hörhold, M., Münch, T., Weißbach, S. *et al.* Modern temperatures in central–north Greenland warmest in past millennium. *Nature* **613**, 503–507 (2023). https://doi.org/10.1038/s41586-022-05517-z

- Received: 08 January 2022
- Accepted: 01 November 2022
- Published: 18 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05517-z

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

Global warming has reached the top of the Greenland ice sheet

Nature Research Briefing 18 Jan 2023

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05517-z

- Article
- Open Access
- Published: 18 January 2023

FinnGen provides genetic insights from a well-phenotyped isolated population

- Mitja I. Kurki,
- Juha Karjalainen,
- Priit Palta,
- Timo P. Sipilä,
- Kati Kristiansson,
- Kati M. Donner,
- Mary P. Reeve,
- Hannele Laivuori,
- Mervi Aavikko,
- Mari A. Kaunisto,
- Anu Loukola,
- Elisa Lahtela,
- Hannele Mattsson,
- Päivi Laiho,
- Pietro Della Briotta Parolo,
- Arto A. Lehisto,
- Masahiro Kanai,
- Nina Mars,
- Joel Rämö,
- Tuomo Kiiskinen,
- Henrike O. Heyne,
- Kumar Veerapen,
- Sina Rüeger,
- Susanna Lemmelä,

- Wei Zhou,
- Sanni Ruotsalainen,
- Kalle Pärn,
- Tero Hiekkalinna,
- Sami Koskelainen,
- Teemu Paajanen,
- Vincent Llorens,
- Javier Gracia-Tabuenca,
- Harri Siirtola,
- Kadri Reis,
- Abdelrahman G. Elnahas,
- Benjamin Sun,
- Christopher N. Foley,
- Katriina Aalto-Setälä,
- Kaur Alasoo,
- Mikko Arvas,
- Kirsi Auro,
- Shameek Biswas,
- Argyro Bizaki-Vallaskangas,
- Olli Carpen,
- Chia-Yen Chen,
- Oluwaseun A. Dada,
- Zhihao Ding,
- Margaret G. Ehm,
- Kari Eklund,
- Martti Färkkilä,
- Hilary Finucane,
- Andrea Ganna,
- Awaisa Ghazal,
- Robert R. Graham,
- Eric M. Green,
- Antti Hakanen,
- Marco Hautalahti,
- Åsa K. Hedman,
- Mikko Hiltunen,
- Reetta Hinttala,
- Iiris Hovatta,

- Xinli Hu,
- Adriana Huertas-Vazquez,
- Laura Huilaja,
- Julie Hunkapiller,
- Howard Jacob,
- Jan-Nygaard Jensen,
- Heikki Joensuu,
- Sally John,
- Valtteri Julkunen,
- Marc Jung,
- Juhani Junttila,
- Kai Kaarniranta,
- Mika Kähönen,
- Risto Kajanne,
- Lila Kallio,
- Reetta Kälviäinen,
- Jaakko Kaprio,
- FinnGen,
- Nurlan Kerimov,
- Johannes Kettunen,
- Elina Kilpeläinen,
- Terhi Kilpi,
- Katherine Klinger,
- Veli-Matti Kosma,
- Teijo Kuopio,
- Venla Kurra,
- Triin Laisk,
- Jari Laukkanen,
- Nathan Lawless,
- Aoxing Liu,
- Simonne Longerich,
- Reedik Mägi,
- Johanna Mäkelä,
- Antti Mäkitie,
- Anders Malarstig,
- Arto Mannermaa,
- Joseph Maranville,

- Athena Matakidou,
- Tuomo Meretoja,
- Sahar V. Mozaffari,
- Mari E. K. Niemi,
- Marianna Niemi,
- Teemu Niiranen,
- Christopher J. O´Donnell,
- Ma´en Obeidat,
- George Okafo,
- Hanna M. Ollila,
- Antti Palomäki,
- Tuula Palotie,
- Jukka Partanen,
- Dirk S. Paul,
- Margit Pelkonen,
- Rion K. Pendergrass,
- Slavé Petrovski,
- Anne Pitkäranta,
- Adam Platt,
- David Pulford,
- Eero Punkka,
- Pirkko Pussinen,
- Neha Raghavan,
- Fedik Rahimov,
- Deepak Rajpal,
- Nicole A. Renaud,
- Bridget Riley-Gillis,
- Rodosthenis Rodosthenous,
- Elmo Saarentaus,
- Aino Salminen,
- Eveliina Salminen,
- Veikko Salomaa,
- Johanna Schleutker,
- Raisa Serpi,
- Huei-yi Shen,
- Richard Siegel,
- Kaisa Silander,

- Sanna Siltanen,
- Sirpa Soini,
- Hilkka Soininen,
- Jae Hoon Sul,
- Ioanna Tachmazidou,
- Kaisa Tasanen,
- Pentti Tienari,
- Sanna Toppila-Salmi,
- Taru Tukiainen,
- Tiinamaija Tuomi,
- Joni A. Turunen,
- Jacob C. Ulirsch,
- Felix Vaura,
- Petri Virolainen,
- Jeffrey Waring,
- Dawn Waterworth,
- Robert Yang,
- Mari Nelis,
- Anu Reigo,
- Andres Metspalu,
- Lili Milani,
- Tõnu Esko,
- Caroline Fox,
- Aki S. Havulinna,
- Markus Perola,
- Samuli Ripatti,
- Anu Jalanko,
- Tarja Laitinen,
- Tomi P. Mäkelä,
- Robert Plenge,
- Mark McCarthy,
- Heiko Runz,
- Mark J. Daly &
- …
- Aarno Palotie

Show authors

Nature **volume 613**, pages 508–518 (2023)

- 7719 Accesses
- 124 Altmetric
- Metrics details

Abstract

Population isolates such as those in Finland benefit genetic research because deleterious alleles are often concentrated on a small number of low-frequency variants (0.1% ≤ minor allele frequency < 5%). These variants survived the founding bottleneck rather than being distributed over a large number of ultrarare variants. Although this effect is well established in Mendelian genetics, its value in common disease genetics is less explored[1,2]. FinnGen aims to study the genome and national health register data of 500,000 Finnish individuals. Given the relatively high median age of participants (63 years) and the substantial fraction of hospital-based recruitment, FinnGen is enriched for disease end points. Here we analyse data from 224,737 participants from FinnGen and study 15 diseases that have previously been investigated in large genome-wide association studies (GWASs). We also include meta-analyses of biobank data from Estonia and the United Kingdom. We identified 30 new associations, primarily low-frequency variants, enriched in the Finnish population. A GWAS of 1,932 diseases also identified 2,733 genome-wide significant associations (893 phenome-wide significant (PWS), $P < 2.6 \times 10^{-11}$) at 2,496 (771 PWS) independent loci with 807 (247 PWS) end points. Among these, fine-mapping implicated 148 (73 PWS) coding variants associated with 83 (42 PWS) end points. Moreover, 91 (47 PWS) had an allele frequency of <5% in non-Finnish European individuals, of which 62 (32 PWS) were enriched by more than twofold in Finland. These findings demonstrate the power of bottlenecked populations to find entry points into the biology of common diseases through low-frequency, high impact variants.

Main

Large biobank studies have become an important source of genetic discoveries. The FinnGen study aims to construct a resource that combines the power of nationwide biobanks, structured national healthcare data and a unique, isolated population. Owing to increased genetic drift, isolated populations with recent bottlenecks can have deleterious, disease-predisposing alleles at considerably higher frequencies than permitted by selection in larger and older outbred populations. Counterbalancing this enrichment of specific low-frequency alleles, the other consequence of a recent bottleneck is that isolated populations have considerably fewer rare variants overall[1,3]. As a result, isolated populations provide an opportunity to identify high-impact disease variants that are rare in other populations[1,2]. In Finland, a strong founding bottleneck occurred about 120 generations ago followed by rapid population expansion. This bottleneck effect has resulted in numerous strongly deleterious alleles that occur more frequently in Finland compared with other European populations. This is manifested in the Finnish Disease Heritage, a set of 36 mostly recessive diseases that are more prevalent in Finland than elsewhere in the world[4]. This population history (which facilitates the identification of low-frequency deleterious alleles) combined with longitudinal information from registers that record hospital in-patient and outpatient diagnoses, purchases of prescription medications and many other national health registries centrally collected for decades provides valuable opportunities for understanding the genetic basis of health and disease.

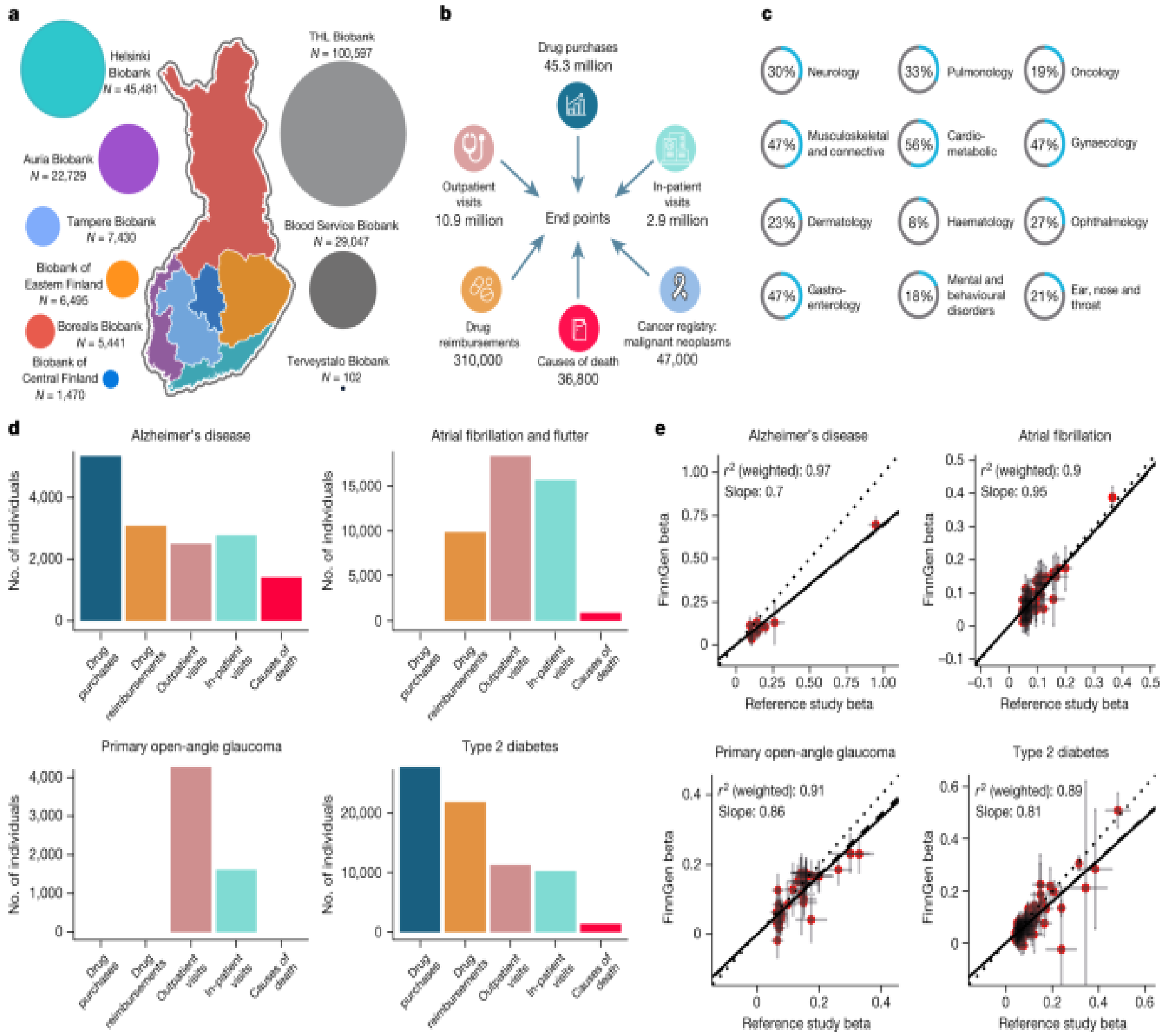
FinnGen is a public–private partnership research project that combines imputed genotype data generated from newly collected and legacy samples from Finnish biobanks and digital health record data from Finnish health registries (https://www.finngen.fi/en) with the aim to provide new insights into disease genetics. FinnGen includes 9 Finnish biobanks, research institutes, universities and university hospitals, 13 international pharmaceutical industry partners and the Finnish Biobank Cooperative (FINBB) in a pre-competitive partnership. As of August 2020 (release 5 described in this article), samples from 412,000 individuals have been collected and have been 224,737 analysed with the aim to have a cohort of 500,000 participants (Supplementary Methods, section 2). The project utilizes data from the nationwide longitudinal health register collected since 1969 from every resident in Finland.

Here we describe the FinnGen project and its current genotype and phenotype content and highlight a series of genetic discoveries from the first data collection phase. In other articles, we describe more detailed studies that showcase different aspects of the rich data available from population registries. Here we first show that FinnGen register-based phenotypes are comparable to those used in disease-specific GWASs in 15 previously well-studied common diseases. We demonstrate the power of the combination of data from an isolated population and other registers to discover new low-frequency variant associations, even in previously well-studied diseases in which FinnGen has a much smaller number of cases than in published disease-specific GWASs. Finally, through a GWAS of 1,932 end points followed by statistical fine-mapping, we demonstrate the ability to identify probable causal coding variants even with low allele frequencies (AFs).

Phenotyping and genotyping

In Finland, similar to the other Nordic countries, there are nationwide electronic health registers that were originally established primarily for administrative purposes to monitor the usage of health care nationwide and over the lifespan of each Finnish resident. These registers have almost complete coverage of major health-related events such as hospitalizations, prescription drug purchases (not including hospital-administered medications), medical procedures or deaths, with a history of data collection spanning more than 50 years. Phenotypes based on health registers (end points) were created by combining data (mainly using classification codes from the International Classification of Diseases (ICD) and the Anatomical Chemical Therapeutic (ACT)) from one or more nationwide health registers (Extended Data Fig. 1, Supplementary Table 1 and Supplementary Figs. 1–4). For the phenome-wide GWAS, we initially constructed more than 2,800 end points by combining data from different health registers, including hospital discharge registers, prescription medication purchase registers and cancer registers (Fig. 1 and Supplementary Methods, section 1; see also https://r5.risteys.finngen.fi/).

Fig. 1: FinnGen sample collection and phenotyping.



a, Samples collected from different geographical areas. The map of Finland is divided into major administrative areas. Coloured regions represent the areas of the nine biobanks that provide samples to FinnGen. The Finnish Institute for Health and Welfare (THL), the Blood Service and the Terveystalo biobanks are not regional. The circle size represents relative sample sizes. The number of samples given are those used in the analyses after QC. **b**, National registries utilized to construct FinnGen end points. The numbers indicate the number of events in each register at the time of FinnGen release. An individual can have multiple diagnoses and can have events from multiple registers contributing to the end point of the individual. **c**, Sample prevalence of major disease categories in FinnGen. Major diseases for each category were chosen for demonstration purposes (Supplementary Tables 3 and 4). **d**, Examples of registers used for constructing four selected end points. The *y* axis represents individuals with

matching register code in each register according to FinnGen end point definitions. Each individual can contribute only once to each register but the same individual can be counted in multiple registers. **e**, Comparison of effect sizes (beta values) in known genome-wide significant loci between four example FinnGen end points and large reference GWAS. The *y* and *x* axes represent FinnGen and reference GWAS beta values respectively. Beta values are aligned to be positive in reference studies. Lines extending from points indicate standard errors of beta values. Regression lines omit intercept and two types of regressions are provided: unweighted and weighted by pooled standard errors from the two studies. Solid line indicates identity line and dotted line and dashed lines indicate unweighted and weighted regression, respectively. Sample sizes used for **e** are given in Supplementary Table 7. Only variants with $P < 1 \times 10^{-10}$ in reference study were included. A comparison of all 15 diseases is provided in the Supplementary Information. Part **a** adapted with permission from an original biobank map created by BBMRI.fi.

FinnGen release 5 presented here contains genotype data for 224,737 individuals after quality control (QC). A total of 154,714 individuals were genotyped with a custom Axiom FinnGen1 array. Data on 70,023 additional individuals were derived from legacy collections (Supplementary Table 2) genotyped with non-custom genotyping arrays (QC details provided in Supplementary Methods, section 3). We developed and utilized a population-specific imputation reference panel of 3,775 high-coverage (25–30 times) whole-genome sequence data for Finnish individuals, containing 16,962,023 single nucleotide polymorphisms, and insertions and deletions (minor allele count of ≥3) (Supplementary Methods, section 3). The majority (16,387,711) of the variants were confidently imputed (information (INFO) score of >0.6; Supplementary Fig. 5).

Population structure and relatedness

To study the genetic ancestry data of 224,737 FinnGen participants that passed genotyping QC (Supplementary Methods, section 3), we combined the FinnGen data with 2,504 phase 3 reference samples from the 1000 Genomes Project[5] and used principal component analysis (PCA) to identify FinnGen participants who have non-Finnish genetic ancestry. Most

participants have broadly Finnish ancestry; 3,676 out of 224,737 (1.63%) outliers were removed (Extended Data Fig. 2 and Supplementary Methods, section 4). We estimated that 165,448 (73.6%) of FinnGen participants have third-degree or closer relatives, which is higher than the estimated 30.3% in the UK Biobank (UKBB)[6]; this result is partially explained by the family-based legacy cohorts in FinnGen. We removed 5,780 duplicates and monozygotic twins (one from each pair removed randomly) and genetic population outliers (Supplementary Methods, section 4) and built a set of approximately unrelated individuals for which the relation between any pair is third degree or higher. In total, we obtained data for 156,977 independent individuals, which were used to compute the PCA, and data for the 61,980 related individuals were projected onto these principal components (PCs) (Supplementary Methods, section 4, and Supplementary Table 5). The first two PCs captured the well-known east–west and north–south genetic differences in Finland[7] (Supplementary Fig. 9). Out of the total remaining 218,957 genotyped samples, we had phenotype data for 218,792 individuals (56.5% females (123,579)), which were then used in all analyses.

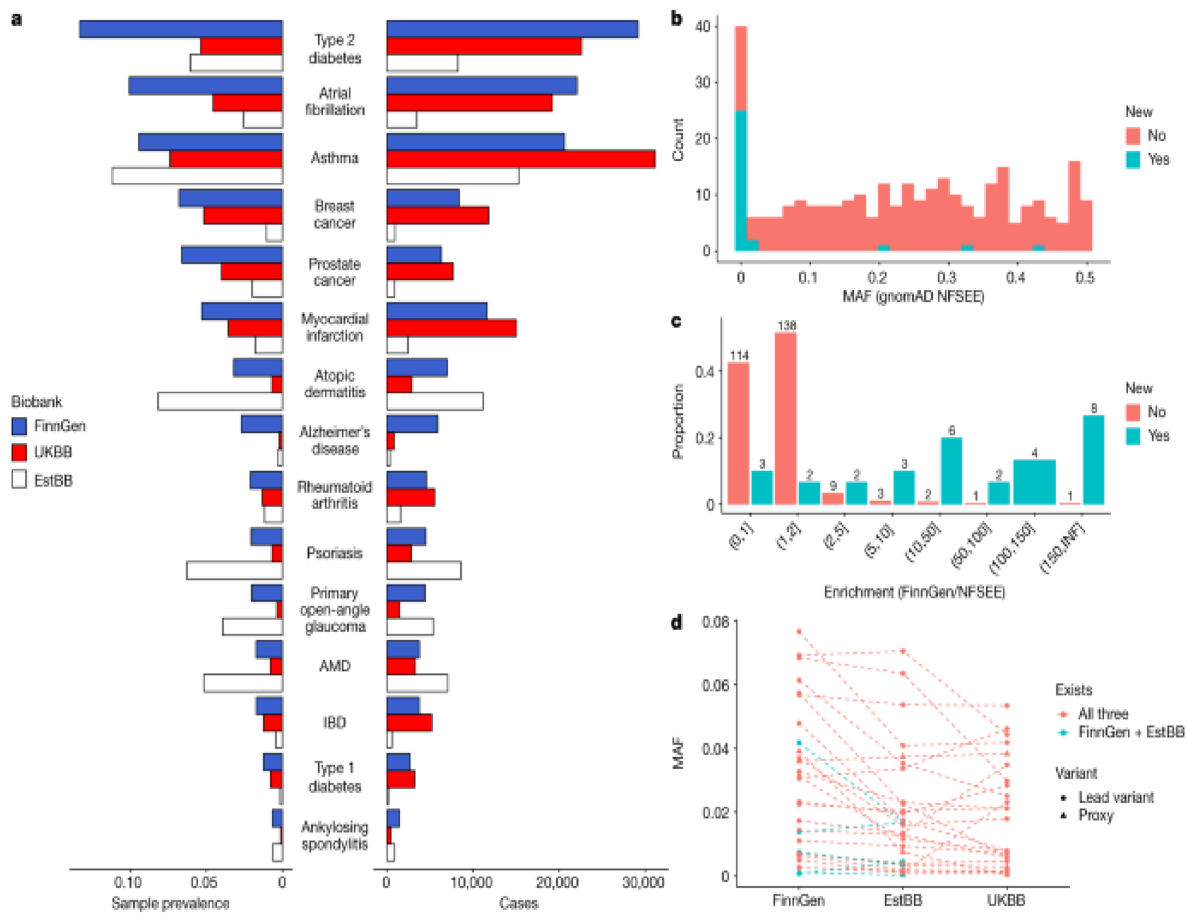
GWAS of nationwide health registries

To benchmark our register-based phenotyping and to explore the value of the isolated setting of Finland, we selected 15 diseases with more than 1,000 cases in FinnGen and for which well-powered GWAS data have been published. We evaluated the accuracy of our phenotyping by comparing the genetic correlations and effect sizes with the previous GWAS results (Supplementary Table 6). None of the genetic correlations were significantly lower than 1 (the lowest genetic correlation was 0.89 (standard error = 0.07) in age-related macular degeneration (AMD); Supplementary Table 6). For diseases with a large number of cases in FinnGen, the effect sizes of lead variants in known loci were largely consistent between FinnGen and previously published meta-analyses. This result demonstrates that our register-based phenotyping is comparable to existing disease-specific GWASs (Fig. 1e, Supplementary Information and Supplementary Table 6). The effect sizes varied more in some diseases that have a smaller number of cases in FinnGen (for example, ankylosing spondylitis, $n = 1462$, $r^2 = 0.62$).

GWAS of these 15 diseases identified 235 loci (that is, regions selected for fine-mapping; Methods) and 275 independent genome-wide significant associations (here onwards, ‘association’ means an independent signal) outside the human leukocyte antigen (HLA) region (GRCh38, chromosome 6: 25–34 Mb). A phenome-wide association study (PheWAS) of FinnGen imputed classical HLA gene alleles has been previously reported[8]. Overall, 44 of the non-HLA associations were driven by low-frequency lead variants (we define ‘low frequency’ as an AF of <5% in non-Finnish, Swedish or Estonian European (NFSEE) individuals in the Genome Aggregation Database (gnomAD; v.2.0.1)[9]) that were more than twice as frequent in Finnish individuals compared with NFSEE individuals. We use NFSEE as a general continental European reference point, excluding individuals from Finland, Sweden and Estonia. As there were large-scale migrations from Finland to Sweden in the twentieth century, many of the chromosomes from sequencing studies of Swedish individuals are of recent Finnish origin. Moreover, the geographically close and linguistically and genetically similar[9] population of Estonia is likely to share elements of the same ancestral founder effect.

Replication of many such enriched variant associations in the Finnish population is hindered by low AFs or missingness in other European populations. People from Finland are genetically more similar to people from Estonia than other European countries[9]. Therefore we first conducted replication using data from 136,724 individuals from the Estonian Biobank (EstBB) and then extended the analysis to individuals from the UKBB (Methods and see Supplementary Table 7 for definitions of end points and case–control numbers). The effect sizes in genome-wide significant hits in FinnGen were mostly concordant with the EstBB (average inverse variance weighted slope of 1.5 (with FinnGen higher) and $r^2 = 0.69$) and the UKBB (slope = 1.1, $r^2 = 0.84$) (Extended Data Fig. 3). FinnGen had a higher case prevalence in the 15 disease diagnoses than in the UKBB, which is probably due to slightly different ascertainment schemes. By contrast, the EstBB had the highest case prevalence in ophthalmic diseases (AMD and glaucoma) and inflammatory skin conditions (atopic dermatitis and psoriasis) (Fig. 2a).

Fig. 2: Comparison of previously unknown and known lead variants in loci identified in the 15 studied diseases.



a, Case prevalence and counts in FinnGen, the EstBB and the UKBB. The phenotypes are sorted on the basis of FinnGen prevalence. **b**, Distribution of minor AFs in known (red) and new (blue) loci in the NFSEE population. **c**, Distribution of AF enrichment between Finland and other Northwestern European populations in gnomAD (excluding Estonia and Sweden). The *x* axis represents enrichment bins. **d**, AFs of 25 replicated genome-wide significant (in FinnGen discovery) new low-frequency (<5% in NFSEE populations) variants in FinnGen, the EstBB and the UKBB. The dotted line indicates the same variants and no line means absence of the variant in other biobanks.

After a meta-analysis of the EstBB and UKBB data, 241 of the 275 associations remained genome-wide significant (Supplementary Table 8). We performed a further meta-analysis of 232 associations that did not meet the genome-wide significance threshold in FinnGen ($5 \times 10^{-8} < P < 1 \times 10^{-6}$), and 57 of those were genome-wide significant after meta-analysis. This meta-analysis resulted in 298 genome-wide significant associations (see

also Supplementary Table 8 for results after multiple testing correction for 15 end points).

To determine whether the observed associations have been previously reported, we queried the GWAS Catalog association database (and largest recent relevant GWAS) for genome-wide significant ($P < 5 \times 10^{-8}$) variants that are in linkage disequilibrium (LD) ($r^2 > 0.1$ in the FinnGen imputation panel) with observed lead variants in FinnGen. As the lowest AF of the new findings was low (0.15%), in addition to published GWASs, we checked whether credible set variants in these loci have also been previously reported in ClinVar. We observed six known pathogenic or likely pathogenic variants, such as a frameshift variant in *PALB2* (p.Leu531fs; AF of 0.1%, not observed outside Finland in gnomAD; Supplementary Table 8) associated with breast cancer. Thirty out of the 298 associations have not been previously reported in the largest published meta-analysis so far (Supplementary Table 6), in a manual literature search, the GWAS Catalog or in ClinVar (Table 1). As expected, we observed that lead variants in novel loci were mostly of low frequency and enriched in Finland compared with known loci from previous GWASs. Specifically, 27 lead variants had minor allele frequency (MAF) values of <5% in gnomAD NFSEE individuals, and 88% of novel and 11% of known loci (after LD pruning, see below) had gnomAD NFSEE MAF values of <5% (Fisher's exact test, $P = 4.29 \times 10^{-17}$). In most cases, the AFs of lower frequency variants (MAF < 5% in gnomAD NFSEE population) were the highest in FinnGen followed by the EstBB and lowest in NFSEE individuals in gnomAD (Fig. 2d).

Table 1 A total of 30 previously unreported associations identified in a GWAS of 15 selected, previously extensively studied phenotypes

Next we performed statistical fine-mapping (Methods) on all 298 genome-wide significant associations (each association is independent; that is, 298 credible sets). Coding variants (missense, frameshift, canonical splice site, stop gained, stop lost or inframe deletion) with posterior inclusion probability (PIP) values of ≥0.05 were observed in 44 (18.7%) out of the 95% credible sets (17 coding variants had PIP > 0.5). Here onwards, we report coding variants with PIP > 0.05 as putatively causal. We recognize that there may be occasions in which assignment of the causal variant to a

coding variant is incorrect (see our accompanying paper[10] for discussions on fine-mapping calibration and replicability). In addition to identifying putative causal coding variants, we sought to identify potential gene expression regulatory mechanisms by colocalizing credible sets with fine-mapped expression quantitative trait locus (eQTL) datasets from the eQTL Catalogue (Methods).

We then wanted to describe the AF spectrum and putative mechanisms of action of risk variants. To do so, we LD pruned the 298 genome-wide significant associations and prioritized the most significant phenotype among the same hits to represent a single putative causal variant (LD r^2 value between lead variants of <0.2). This process resulted in 281 previously unknown associations (27 new).

Most of the 281 previously unknown associations were common variant associations. However, 53 of these had a lead variant frequency of less than 5% in NFSEE individuals, and 38 of them were enriched by more than two times in the Finnish population compared with the NFSEE population. We observed a coding variant more often in the credible sets of associations that were enriched by more than twofold (19 out of 38; 50%) than in non-enriched associations (6 out of 15; 40%) at lower frequencies (MAF < 5%).

Following the discovery of 27 new associations, we sought to determine potential mechanisms of action through the identification of coding variants in their credible sets and potential regulatory effects by colocalization with eQTL associations from the eQTL Catalogue. We identified putative causal coding variants in 9 out of 27 loci and eQTL colocalization in 4 out of 27 loci. In two out of the four eQTL loci, we observed a coding variant in credible sets (*IL4R* and *MYH14*; the eQTLs point to different genes than the coding variants). The two remaining eQTL colocalizations were breast cancer loci colocalizing with *H2BP2* eQTL in lung tissue and type 2 diabetes colocalizing with *PRRG4* in lipopolysaccharide-stimulated monocytes. The disease relevance of these eQTLs is currently not evident.

No credible coding variants or eQTLs were identified in 16 out of 27 loci (Supplementary Table 8). The fraction of associations in which we observed eQTLs was small (14.8%). Most of the new associations were driven by variants with low AFs in NFSEE populations (Table 1 and Fig. 2b,d). The

low fraction of observed eQTL colocalizations is probably explained by the low AF of 25 out of the 27 of the variants in available eQTL studies (such as GTEx), for which the majority of the samples do not have Finnish or Estonian ancestry.

We next aimed to explore the benefits of the FinnGen dataset in GWAS discovery. We extrapolated observed meta-analysis results in FinnGen, the UKBB and the EstBB to match the sample size of the UKBB in 14 demonstration diseases (excluding Alzheimer's disease; Supplementary Methods). The distribution of extrapolated *P* values was shifted towards greater significance in FinnGen compared with those of the UKBB and the EstBB in a matched total sample size scenario for the 14 demonstration diseases (Supplementary Methods and Supplementary Fig. 11). Moreover, frequency enrichment was a major driver in the gain of power in low-frequency variants (Supplementary Fig. 12). In individual end points with similar sample prevalence in FinnGen and the UKBB, similar for inflammatory bowel disease (IBD), the greatest gain in power was in variants in which the AFs are <0.5% in the UKBB (see Supplementary Fig. 13 for a comparison for each end point and biobank).

The identification of a new signal for IBD mapping to a single variant in an intron of *TNRC18* highlights the value of FinnGen for discovery, even when the case sample size is below that of existing meta-analyses. This variant has a strong risk-increasing effect (AF = 3.6%, odds ratio (OR) = 3.2, $P = 2.4 \times 10^{-61}$), which eclipses the significance of signals at *IL23R*, *NOD2* and the major histocompatibility complex. The variant is enriched by 114-fold in the Finnish population compared with the NFSEE population, in whom the AF is too low (0.04%) to have been identified in previous GWASs (this FinnGen association was also reported in ref. [11]). We were, however, able to replicate this association in the EstBB (AF = 1.3%, OR = 3.9, $P = 2.8 \times 10^{-6}$) owing to the relatively higher frequency in the genetically related Estonian population. This variant was also associated with risk for multiple other inflammatory conditions evaluated in FinnGen, including interstitial lung disease (OR = 1.43, $P = 6.3 \times 10^{-26}$), ankylosing spondylitis (OR = 4.2, $P = 1.8 \times 10^{-34}$), iridocyclitis (OR = 2.3, $P = 1.2 \times 10^{-27}$) and psoriasis (OR = 1.6, $P = 1.1 \times 10^{-13}$). However, the same allele appears to be protective for an end point that combines multiple autoimmune diseases

(https://r5.risteys.finngen.fi/phenocode/AUTOIMMUNE) (OR = 0.84, $P = 6.2 \times 10^{-12}$; for example, type 1 diabetes (OR = 0.64, $P = 2.7 \times 10^{-7}$) and hypothyroidism (OR = 0.85, $P = 7.8 \times 10^{-7}$).

The highest number (eight loci) of new and enriched low-frequency associations were identified in type 2 diabetes, which is probably due to the large number of patients with type 2 diabetes in FinnGen release 5 (29,193). Other noteworthy observations from this set of 30 findings for 15 well-studied diseases are described in Supplementary Note 1.

Coding variant associations

Motivated by the identification of high-effect coding variant associations within the selected 15 diseases, we performed a PheWAS followed by fine-mapping to identify putative causal coding variants enriched in the Finnish population.

In a GWAS of 1,932 distinct end points and 16,387,711 variants (Supplementary Table 4; case overlap < 50% and *n* cases > 80), we identified 2,733 independent associations in 2,496 loci across 807 end points (Supplementary Table 9) at a genome-wide significance threshold ($P < 5 \times 10^{-8}$). Moreover, 893 signals in 771 loci across 247 end points at PWS thresholds ($P < 2.6 \times 10^{-11}$) were identified. The HLA region was excluded here, and a PheWAS of imputed classical HLA gene alleles in FinnGen is reported in ref. [8].

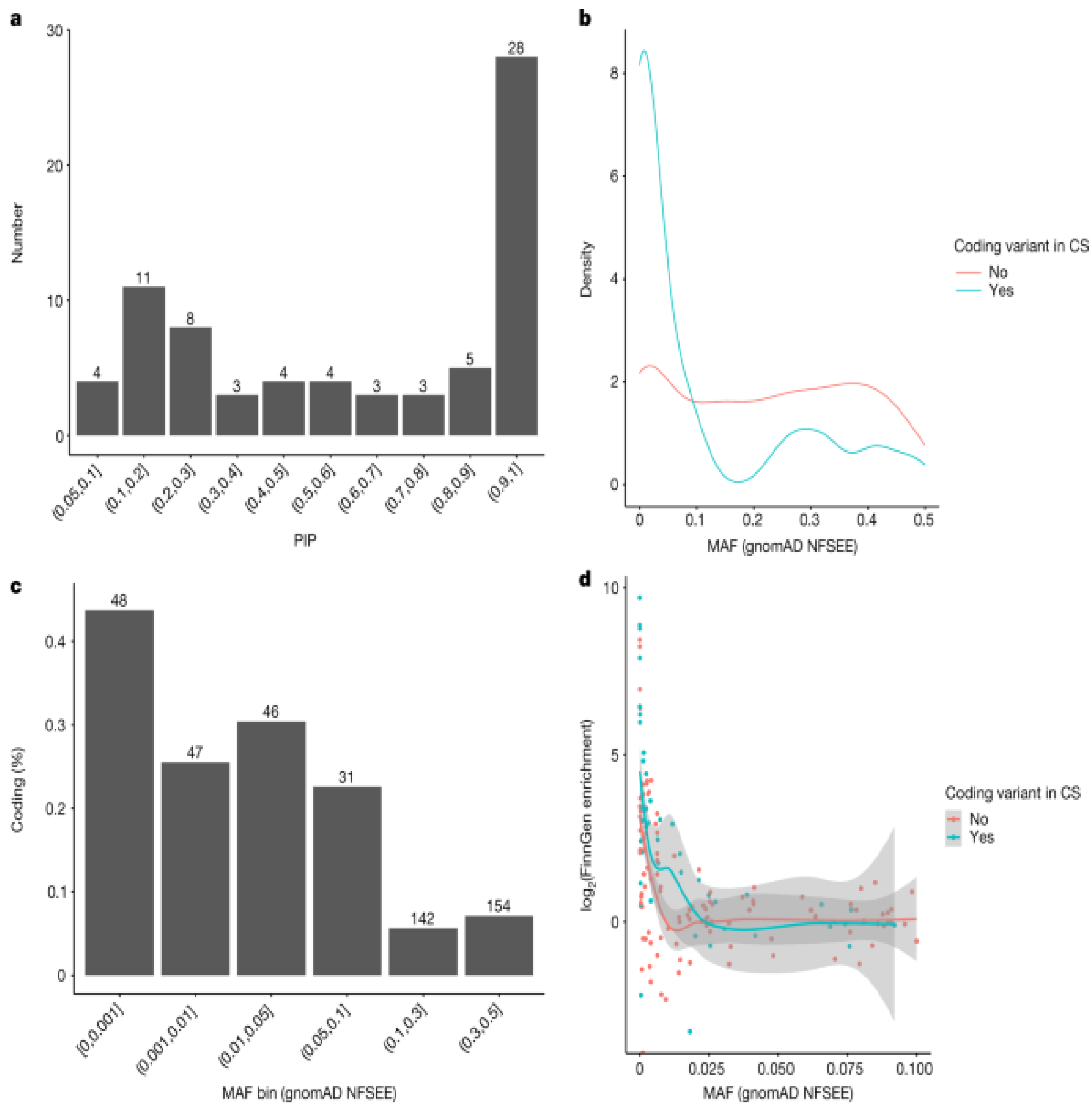
Using statistical fine-mapping, we observed a coding variant (missense, frameshift, canonical splice site, stop gained, stop lost or inframe deletion; PIP > 0.05) in 369 associations (13.5% of all associations) spanning 202 end points. Full results with all 2,803 end points (including end points with a case overlap of >50% that are excluded here) are publicly available from a customized browser based on the PheWeb code base (https://r5.finngen.fi) and as summary statistic files (https://www.finngen.fi/en/access_results).

To put the frequency spectrum and putative mechanisms of action in an interpretable context, we chose a single most-significant association per signal by LD-based merging ($r^2 > 0.3$ lead variants merged), which resulted

in 1,838 unique associations in 681 end points (Supplementary Table 10). Overall, 493 of the associations in 112 end points were PWS ($P < 2.6 \times 10^{-11}$). Although most of the 493 PWS unique associations were driven by common variants, 143 and 97 had a lead variant frequency of <5% and <1%, respectively, in gnomAD NFSEE populations. We observed that 82 (57.3%) of the 143 low-frequency (MAF < 5%) lead variants were enriched by more than twofold in Finland compared with NFSEE populations. To estimate the number of putative new associations, we searched for known significant associations using the Open Targets API platform (GWAS Catalogue and the UKBB) and ClinVar for each of the 1,838 associations. Among these, 864 (47%) were not associated with any phenotype in those databases (75 out of 493 (15%) of the stringent $P < 2.6 \times 10^{-11}$ associations). The fraction of previously unreported associations among genome-wide significant (702 out of 841 (84%)) and stringent (69 out of 143 (48%)) associations were notably higher among low-frequency variants (MAF < 5% in NFSEE individuals).

After statistical fine-mapping of the 493 unique PWS associations, we identified a coding variant (PIP > 0.05) in 73 (14.8%) of the credible sets associated with 42 end points (Supplementary Table 10). Most (43) of the fine-mapped coding variants had PIP values of >0.5 and 28 had PIP values of >0.9 (Fig. 3a). The highest proportion and the majority (54 out of 73) of associated coding variants had NFSEE MAF < 10% (Fig. 3b,c). The coding variant associations were more enriched in Finland than noncoding associations in associations driven by variants with AFs of <5% in NFSEE people (Fig. 3d; Wilcoxon rank sum test $P = 3.6 \times 10^{-3}$). For example, we observed a coding variant in 42% (34 out of 89) of the associations with a lead variant that was enriched by more than two times in Finland compared with NFSEE people among low-frequency associations (NFSEE MAF < 5%). By contrast, the proportion of coding variants was lower at 21.7% (13 out of 60) in non-enriched associations (see Extended Data Fig. 4 for enrichment in various NFSEE MAF bins). The higher proportion of coding variants in those that were enriched by more than two times persisted when the PIP threshold was increased to 0.2 (enriched, 30 out of 77 (35.8%); non-enriched, 11 out of 58 (18.9%)).

Fig. 3: Characteristics of unique associations in end points identified in FinnGen.



Characteristics of 493 (73 with coding variants in the credible set) specific associations in 112 (42 end points with coding variants in the credible set) end points identified in FinnGen release 5. Note that 25 of the associations with a coding variant with PIP < 0.05 in credible sets were removed from plots as ‘uncertain to contain coding variant’. **a**, Distribution of fine-mapping PIP values of the 73 coding variants. **b**, AF spectrum in associations with and without coding variants in credible sets (CS). **c**, Proportion of coding variants identified in different AFs (in NFSEE individuals in gnomAD). The numbers above the bars indicate the number of associations within a bin, the *y* axis indicates the proportion of associations

with coding variants in their credible sets. **d**, Enrichment in Finland as a function of AF in the gnomAD NFSEE population (enrichment value for variants with AF values of 0 in NFEE individuals in gnomAD was set to maximum observed enrichment value of $\log_2(166) = 7.38$). The smoothed regression lines of local average enrichment are estimated by local polynomial fitting (loess) and the shaded areas represent 95% confidence intervals of the model fit.

The fine-mapping properties and replicability of 67 FinnGen traits across diverse biobanks (FinnGen, Biobank Japan and the UKBB) are explored in detail in another manuscript[10], and functional variant associations in the UKBB and FinnGen are described in ref. [12].

We next wanted to quantify the benefits of population isolates such as Finland in GWAS discovery. To this end, we assessed whether lower frequency (MAF < 5% in NFSEE people) variants enriched in the Finnish population were more likely to be associated with a phenotype than would be expected by chance. We randomly sampled 1,000,000 times the number of genome-wide significant variants observed (143) from a set of frequency-matched variants (MAF NFSEE < 5%) that were not associated with any end point ($P > 0.001$). None of the 1 million random draws had a higher proportion of variants enriched by more than twofold in the Finnish population than was observed in the significant associations (57.3% observed versus 33% expected; $P = 1.0 \times 10^{-16}$).

Known pathogenic variant associations

Among the genome-wide significant coding variant associations, we identified 13 variant associations (AF range of 0.04–2%) classified as pathogenic or likely pathogenic in ClinVar (Supplementary Table 10). Nine out of the 13 variants were enriched by more than 20-fold in Finland compared with NFSEE populations. Some of these variants have previously been primarily considered recessive. Here, however, we observed that some were a risk variant in the heterozygous state. An example is a rare frameshift variant at *NPHS1* associated with nephrotic syndrome, including the congenital form (ICD-10: N04,p.Leu41fs; AF FinnGen = 0.9%; gnomAD NFSEE = 0.009%; OR = 185, $P = 4.3 \times 10^{-27}$). Congenital nephrotic

syndrome in Finnish individuals is a recessively inherited rare disease, and is in the Finnish Disease Heritage database[4]. The pathogenic variant associations listed in ClinVar include a missense variant in *XPA* (xeroderma pigmentosum) associated with non-melanoma neoplasm of skin ('other malignant neoplasm of skin') (p.Arg228Ter; AF FinnGen = 0.02%, gnomAD NFSEE = 0%; OR = 4.4, $P = 8.3 \times 10^{-18}$), and the abovementioned frameshift variant in *PALB2* associated with breast cancer (p.Leu531fs, 'malignant neoplasm of breast'; p.Ala82Pro; AF FinnGen = 0.2%, gnomAD NFSEE = 0%; OR = 28.8, $P = 3.7 \times 10^{-33}$). Furthermore, a known pathogenic recessively acting missense variant in *CERKL* was associated with hereditary retinal dystrophy (p.Cys125Trp; AF FinnGen = 0.6%, gnomAD NFSEE = 0%; OR = 98,716, $P = 5.15 \times 10^{-25}$). This association is, however, driven by compound heterozygotes, as previously detailed[13]. These associations demonstrate that imputation using a population-specific genotyping array and an imputation panel combined with national-registry-based phenotyping in the isolated Finnish population can successfully identify associations and fine-map causal variants even in rare variants and phenotypes. An extended study of ClinVar variants and variants with specific biallelic Mendelian effects in FinnGen is provided in a companion paper[13].

Associations in known disease genes

In the remaining 135 genome-wide significant coding variant associations not reported as pathogenic in ClinVar, 77 had NFSEE MAF values of <5%. Of the 77 variants, 54 were more than 5 times more common in Finland than in NFSEE populations, and 19 had not been previously observed in NFSEE people (Supplementary Table 2). Nine out of the 19 variants are in a gene in which other variants are pathogenic for various traits, 3 of which are for the same or related traits. These FinnGen associations include the following variants: a *RFX6* frameshift variant associated with type 2 diabetes (p.His293LeufsTer7; AF = 0.15%, OR = 3.7, $P = 1.2 \times 10^{-10}$; ClinVar, 'monogenic diabetes and others'); a *TERT* missense variant (AF = 0.15%, OR = 1,032, $P = 6.5 \times 10^{-21}$) associated with idiopathic pulmonary fibrosis (ClinVar, 'idiopathic pulmonary fibrosis'); a missense in *MYH14* associated with sensorineural hearing loss (p.Ala1156Ser; AF = 0.04%, OR = 19.9, $P = 1 \times 10^{-15}$; ClinVar, 'non-syndromic hearing loss' and others); and a stop

gained variant in *TG* associated with autoimmune hypothyroidism (p.Gln655Ter; AF = 0.1%, OR = 3.2, $P = 3.9 \times 10^{-11}$). These variants in *RFX6, TERT* and *TG* have been previously observed in Finnish and Nordic cohorts[14,15,16], but had uncertain significance (single carrier in *TG*) or conflicting interpretation (*TERT*) in ClinVar. Pathogenic variants in *RFX6* cause Mitchell–Riley syndrome with recessive inheritance (characterized by neonatal diabetes). However, heterozygote enrichment of *RFX6*-truncating variants have been observed in maturity-onset diabetes of the young[14], for which the same variant observed here was identified in a replication in Finnish data. RFX6 is a regulator of transcription factors involved in beta-cell maturation and has a specific role in releasing gastric inhibitory peptide (GIP) and GLP1 in response to meals. Our results propose that around 1:700 individuals in Finland carry a frameshift variant that has been previously shown to reduce incretin levels and to lead to isolated diabetes[14]. It is tempting to speculate that early administration of GLP1 analogues would benefit carriers of this diabetes-associated variant.

New disease associations

Among the previously undescribed genome-wide significant coding variant associations without previous associations in Open Targets (GWAS Catalog and the UKBB) or ClinVar, we observed 29 that had NFSEE MAF values of <5% and were 2 times more frequent in Finland, 9 of which had no copies in NFSEE populations (Supplementary Table 11). We summarize selected new discoveries and biological knowledge gained in Supplementary Table 12. A missense variant not observed outside Finland (p.Val70Phe; AF = 0.2%, OR = 3.0, $P = 2.1 \times 10^{-9}$) in *PLTP* was associated with coronary revascularization (*n* = 12,271 coronary angioplasty or bypass grafting). *PLTP* is a lipid-transfer protein in human plasma that transfers phospholipids from triglyceride-rich lipoproteins to high-density lipoprotein, and its activity is associated with atherogenesis in humans and mice[17]. Noncoding variations near *PLTP* independent of p.Val70Phe are associated with lipid levels (high-density lipoprotein and triglycerides)[18] and coronary artery disease[19]. The identification of a coding variant in this gene provides support for *PLTP* as the causal gene for symptomatic atherosclerosis in this locus. Other variants associated with coronary artery disease included a missense variant (p.Gly567Arg; AF = 0.9%, OR = 2.0, $P = 5.2 \times 10^{-12}$) in

HHIPL1, which was associated with coronary revascularization (n = 12,271), and a splice acceptor variant (c.7325-2A>G; AF = 0.7%, OR = 2.5, $P = 2.9 \times 10^{-08}$) in *NBEAL1*, which was associated with coronary artery bypass grafting (n = 5,779). Both genes are susceptibility loci for coronary artery disease[19] and have been suggested as causal, although for *NBEAL1* the evidence is inconsistent[20]. *HHIPL1* encodes a secreted sonic hedgehog regulator that modulates atherosclerosis-relevant smooth muscle cell phenotypes and promotes atherosclerosis in mice[21]. *NBEAL1* regulates cholesterol metabolism by modulating low-density lipoprotein (LDL) receptor expression, and genetic variants in *NBEAL1* are associated with decreased expression of *NBEAL1* in arteries[22]. Our results strengthen the evidence that both these genes are causal in the loci.

A missense variant in *LAG3* (p.Pro67Thr; AF = 0.08%, gnomAD NFSEE = 0%) was associated with autoimmune hypothyroidism (n = 22,997, OR = 3.2, $P = 4.6 \times 10^{-8}$, lead variant $P = 4.57 \times 10^{-8}$). *LAG3* encodes an immune checkpoint protein that is involved in inhibitory signalling of immune response, especially in T cells[23]. LAG3 has been a target of active immune checkpoint inhibitor cancer immunotherapy development. One such immunotherapy was recently approved by the US Food and Drug Administration as a combination treatment for unresectable or metastatic melanoma[24]. Immune checkpoint inhibition therapies aim to enhance immune responses against tumour cells. Excessive immune responses, however, can exert deleterious effects on healthy tissue and lead to autoimmune disease. A common side effect of immune checkpoint inhibitors, including those that target LAG3, is hypothyroidism. The p.Pro67Thr variant could be acting as an inhibitor of LAG3 immunoregulatory activity, which in turn leads to susceptibility to hypothyroidism. In a PheWAS of p.Pro67Thr, we observed a nominally increased risk for other immune-related conditions (for example, psoriatic arthropathies (M13_PSORIARTH_ICD10) n = 1,455, OR = 7.8, $P = 3.3 \times 10^{-3}$; urticaria and erythema (L12_URTICARIAERYTHEMA), n = 6,328, OR = 3.7, $P = 2.7 \times 10^{-4}$; and streptococcal septicaemia (AB1_STREPTO_SEPSIS), n = 1,090, OR = 15, $P = 2.2 \times 10^{-3}$), but we did not observe protective effects with any cancers. It should be noted, however, that owing to the rarity of the variant, the data were not sufficiently powered to detect more subtle effects.

We found a missense variant (p.Tyr212Phe, rs35937944) in *COLGALT2* that was enriched by >20-fold in the Finnish population. This variant was associated with a reduced risk for arthrosis (OR = 0.79, $P = 2.57 \times 10^{-10}$), coxarthrosis (OR = 0.68, $P = 1.34 \times 10^{-19}$) and gonarthrosis (OR = 0.80, $P = 7.5 \times 10^{-7}$). A noncoding variant near *COLGALT2* has recently been described as a GWAS locus for osteoarthritis[25]. *COLGALT2* encodes the procollagen galactosyltransferase 2, which initiates post-translational modification of collagens by transferring β-galactose to hydroxylysine residues, an important step to ensure structure and function of bone and connective tissue. Modulating COLGALT2 enzymatic activity with drugs could be a potential strategy to reduce arthritis risk.

CD63 is a cell surface protein involved in basophil activation and mast cell degranulation. We identified a missense variant in *CD63* (rs148781286) that was enriched by >42-fold in the Finnish population. This variant was associated with childhood asthma (OR = 3.5, $P = 3.37 \times 10^{-9}$). In a combined analysis with data from the EstBB and the UKBB, this variant was also associated with atopic dermatitis[26]. Mediators secreted by basophils and mast cells correlate with asthma severity in the clinic, and a CD63-based basophil activation test has been reported to predict asthma outcome in young children with wheezing episodes[27]. The observation of a putative causal relationship between genetic variations in *CD36*, basophil activation and childhood asthma risk and severity may point to a new intervention point for targeted asthma therapies.

A missense variant in *TUBA1*C (p.Ala331Val; AF = 0.2%, OR = 35.2, $P = 1.4 \times 10^{-10}$) was associated with sudden idiopathic hearing loss (*n* = 1,491). No relevant phenotype has previously been reported for variants in *TUBA1C*. *TUBA1C* encodes an α-tubulin isotype. The precise roles of α-tubulin isotypes are unknown, but mutations in other tubulins can cause various neurodevelopmental disorders[28]. The p.Ala331Val variant was also associated with vestibular neuritis (inflammation of the vestibular nerve; *n* = 1,224, OR = 40.9, $P = 3.2 \times 10^{-10}$). Pure vestibular neuritis presents acutely with vertigo but not hearing loss, and accurate diagnosis of vertigo in acute settings is challenging and misdiagnosis is possible.

A >30-fold-enriched missense variant, pThr155Met (rs145955907), in *ZAP70* was associated with sarcoidosis (OR = 2.05, $P = 1.03 \times 10^{-8}$). Previously, homozygote or compound heterozygote mutations in *ZAP70* have been described in cell-mediated combined immunodeficiency caused by abnormal T cell receptor signalling[29]. Associations of heterozygote variants have not been associated with any disease so far. Given its crucial role in cell signalling, the ZAP70 association with sarcoidosis seems in line with its key role in immunity.

A 75-fold-enriched missense variant, p.Ala777Thr (rs199680517), in *PPP1R26* was associated with endometriosis (OR = 1.97, $P = 3.41 \times 10^{-8}$). PPP1R26 (protein phosphatase 1 regulatory subunit 26) has been associated with tumour formation and has been observed to be upregulated in various malignancies. Cellular GWAS analyses have identified one variant to be associated with carboplatin-induced toxicity[30]. In one study, a copy number variant has been associated with endometriosis, but how this gene contributes to endometriosis susceptibility remains speculative[31].

We also report several of these coding associations in separate manuscripts. One such new observation is a missense variant (p.Arg20Gln; AF = 3%, gnomAD NFSEE = 0.7%) in *SPDL1* with a pleiotropic association. It is associated with a strongly increased risk of idiopathic pulmonary fibrosis (OR = 3.1, $P = 1.0 \times 10^{-15}$) but protective with an end point that combines all cancers (OR = 0.82, $P = 2.1 \times 10^{-15}$)[32]. Other associations between variants and disease described in separate manuscripts include the following: an inframe deletion in *MFGE8* and coronary atherosclerosis (p.Asn239dup; AF = 2.9%, gnomAD NFSEE = 0%, OR = 0.74, $P = 5.4 \times 10^{-15}$)[33]; a frameshift variant in *MEPE* (p.Lys101IlefsTer26; AF = 0.3%, gnomAD NFSEE = 0.07%, OR = 18.9, $P = 1.5 \times 10^{-11}$) and otosclerosis[34]; and a missense variant in *ANGPTL7* (p.Arg220Cys; AF = 4.2%, gnomAD NFSEE = 0.06%, OR = 0.7, $P = 7.2 \times 10^{-16}$) and glaucoma[35].

Coding variants associated with drug use

An notable registry available in FinnGen is a prescription medication purchase registry (KELA; Supplementary Table 1), which links all prescription medication purchases for all FinnGen participants since 1995.

Using prescription records from this registry, we identified two enriched low-frequency coding variants that were associated with drug purchase of statin medications (three or more purchases per individual) (Supplementary Table 11). A missense variant in *TM6SF2* (p.Leu156Pro, rs187429064) was associated with a decreased likelihood of being prescribed statins (AF = 5.2%, gnomAD NFSEE = 1.2%; OR = 0.86, $P = 3.8 \times 10^{-13}$) but with an increased likelihood for insulin medication for diabetes (OR = 1.17, $P = 8.2 \times 10^{-11}$) and type 2 diabetes (OR = 1.15, $P = 2.6 \times 10^{-8}$). In addition, the same variant showed a strong association with a strongly increased risk of hepatocellular carcinoma (ICD-10 C22 'hepatic and bile duct cancer'; OR = 3.7, $P = 5.9 \times 10^{-10}$). The hepatic and bile duct cancer association did not change after conditioning on statin medication (OR = 3.7, $P = 7.1 \times 10^{-10}$). Consistent with a decrease in the likelihood of being prescribed statins, *TM6SF2* p.Leu156Pro and another independent ($r^2 = 0.003$) missense variant (p.Gly167Lys, rs58542926) have previously been associated with decreased LDL and total cholesterol levels[36]. In a mouse model, both p.Gly167Lys and Leu156Pro lead to increased protein turnover and reduced cellular TM6SF2 levels[37]. *TM6SF2* p.Gly167Lys leads to decreases in hepatic large, very LDL particle secretion and increases in intracellular lipid accumulation[38]. These effects probably explain its associations with non-alcoholic fatty liver disease[39], alcohol-related cirrhosis[40], hepatocellular carcinoma[41] and incident type 2 diabetes[42]. Our results provide, in a single PheWAS analysis, strong evidence of a previously unknown p.Leu156Pro variant that has similar consequences of decreasing circulating lipid levels and increasing the risk of diabetes, cirrhosis and liver cancer, as observed for p.Gly167Lys. Such pleiotropy of the variant can be explored in the custom PheWeb browser (http://r5.finngen.fi/variant/19-19269704-A-G).

Conclusions

In this paper and accompanying publications, we present FinnGen, one of the largest nationwide genetic studies with access to comprehensive electronic health register data of all participants. The final aim of the study is to collect data for 500,000 biobank participants by the end of 2023. The interim releases of FinnGen have already contributed to many new discoveries and insights into human genetic variation and how it affects

disease and health[35,43,44,45,46,47], including contributions to the COVID-19 host genetics initiative[48] and the global biobank meta-analysis initiative[49]. Summary statistics from each data release will be made publicly available after a 1-year embargo period, and all summary statistics described here are freely available at www.finngen.fi/en/access_results.

An important feature of FinnGen compared with other similar projects, such as the UKBB[6], is the specific genetic makeup of the Finnish population. In the GWAS of selected, well-studied diseases, we were able to identify several new associations with a fraction of the cases compared with the largest published GWAS. These associations were largely observed with variants that were increased in frequency in the Finnish population bottleneck and would have required prohibitively large sample sizes in older, non-bottlenecked populations (Fig. 2d).

Moreover, in the GWAS of 1,932 end points, we observed that variants in the Finnish population that were enriched by more than twofold were 1.7-times more likely to be associated with a phenotype than would be expected by chance.

Furthermore, we observed that putative coding variant associations were not only of lower AF but also more often enriched in Finland than noncoding variant associations (Fig. 3). This observation is expected, as coding variant associations are more deleterious on average and selection drives the AFs down. However, some of these deleterious alleles survived the bottleneck and increased in frequency, which facilitated the identification of their associations with diseases.

Imputation with a population-specific imputation panel provides high imputation accuracy down to very low AFs (Supplementary Fig. 5), which enabled the identification of associations with low-frequency variants using a GWAS approach instead of direct sequencing. This high imputation accuracy combined with broad population registry-based phenotyping facilitates the identification of very low-frequency variants associated with rare phenotypes, which have largely been missed in the majority of GWASs published so far[50]. We demonstrated this by identifying known ClinVar variant associations with diseases such as congenital nephrotic syndrome or polycystic liver disease, which are both registered in the Finnish Disease

Heritage database. Furthermore, we uncovered new low-frequency variant associations with common and rare phenotypes, including clinically challenging but not well genetically studied sudden idiopathic hearing loss or carpal tunnel syndrome. The recently reported[35] Gln175His variant in *ANGPT7*, which is enriched in the Finnish population and is protective against glaucoma, is also an example of the benefit of the bottleneck effect in the discovery of disease-associated variants.

The university-hospital-based recruitment, together with legacy case cohorts of several diseases, is another feature of FinnGen. This strategy captures cases in many disease areas and distinguishes it from many working-age population cohorts. For example, in the UKBB, in which recruitment was based on postal invitation to individuals aged 40–69 years and living within 40 km (25 miles) of one of the assessment centres[51], the participants are likely to be healthier than in hospital-based collections. The approach in FinnGen has advantages and disadvantages. For many disease-focused studies, it provides a higher number of cases and a relatively economical way of recruiting a large sample within a feasible time frame. For example, in the 15 common diseases studied in this paper, the sample prevalence in FinnGen was higher than in the UKBB. The difference was the most extreme for Alzheimer's disease (2.7% in FinnGen compared with 0.2% in UKBB), a disease of old age, and the most similar in asthma (9.4% in FinnGen compared with. 7.4% in the UKBB) (Fig. 2a). FinnGen also has a relatively high sample prevalence of severe mental disorders such as schizophrenia (2.5%, $n = 5,562$) and bipolar disease (2.1%, $n = 4,501$), which are often underrepresented in biobank studies. A key aspect of the recruitment strategy for the Finnish biobank is that legislation enables participants to donate samples with broad consent to medical research in general. This makes recruitment cost-effective, as the same samples and data can be used, after appropriate application steps, for many medical research studies. However, owing to the recruitment strategy, FinnGen is not epidemiologically representative, and some disease prevalence estimates might be over or underrepresented in FinnGen compared with population values (for example, asthma is 10.4% in FinnGen, 7.7 in FinRegistry, and type 2 diabetes is 14.5% in FinnGen, 8.2% in FinRegistry (https://www.finregistry.fi/)). The recruitment strategies for FinnGen are not anticipated to cause significant biases to the GWAS results presented here,

but would be an aspect to consider, for example, when studying disease progression or building predictive models. We further explored the benefit of the FinnGen approach and showed that data from FinnGen has greater discovery power than data from the UKBB in a matched sample size scenario for 14 common diseases (Supplementary Fig. 11).

In conclusion, FinnGen as a large-scale biobank resource with specific features of the Nordic healthcare system and population structure provides opportunities for a wide range of genetic discoveries. These include identification of disease-associated coding variants, identification of variant pleiotropy and longitudinal analyses of disease trajectories. Combining results with other large-scale biobank projects can further improve our understanding of the role of genetic variation in health and disease, especially in genetically understudied diseases.

Methods

Biobank samples

The FinnGen study (https://www.finngen.fi/en) is an ongoing research project that utilizes samples from a nationwide network of Finnish biobanks and digital healthcare data from national health registers. FinnGen aims to produce genomic data with linkage to health register data of 500,000 biobank participants. Samples in the FinnGen study include legacy samples (prospected number 200,000) from previous research cohorts (often disease-specific) that have been transferred to the Finnish biobanks, and prospective samples (prospected number 300,000) collected by biobanks across Finland. Prospective samples from six regional hospital biobanks represent a wide variety of patients enrolled in specialized health care, samples from a private healthcare biobank enable enrichment of the FinnGen cohort with patients underrepresented in specialized health care, whereas participants recruited through the Blood Service Biobank enrich the cohort with healthier individuals. Samples have not specifically been collected for FinnGen, but the study has incorporated all that have been available in the biobanks (see Supplementary Methods for details). In the current study, we included samples from 224,737 biobank participants.

Phenotyping

Registry data on all FinnGen participants were collected and processed from the following different national health registers: hospital and outpatient visits in HILMO, a care register for health care (in-patient and outpatient primary and secondary diagnoses: ICD-8, ICD-9 and ICD-10; operations: NOMESCO Classification of Surgical Procedures and Hospital League surgical procedure codes); AvoHILMO, a register of primary health care (main and secondary diagnosis using ICD-10 and ICPC2 codes, operations and procedures using NOMESCO and national SPAT codes); Cause of Death (immediate, underlying and contributing causes of death on the death certificate with ICD-8, ICD-9 and ICD-10 codes); reimbursed medication entitlements and prescribed medicine purchases (specific Social Insurance Institution of Finland reimbursement codes and ATC codes, respectively); and the Finnish Cancer Registry (using ICD-O-3 codes). Pseudonymized register data were combined with the minimum phenotype dataset from the Finnish biobanks (age, sex, year of sampling, height, weight and smoking status). Clinical end points were constructed from the register codes using the Finnish version of the International Classification of Diseases, 10th revision (ICD-10) diagnosis codes and harmonizing those with definitions from ICD-8 and ICD-9. The Finnish ICD version is mostly identical to the international ICD classification, but has minor modifications. For example, there are additions to certain disease classifications in the fourth and fifth character level to add specificity. When relevant, the information on reimbursed medication and/or prescription medicine purchases and operations augmented the end point data. Cancer end points were constructed on the basis of the Finnish Cancer Registry and Cause of Death data. The definitions of FinnGen disease end points and their respective controls for each release are available at https://www.finngen.fi/en/researchers/clinical-endpoints, and FinnGen end points can also be browsed at https://r5.risteys.finngen.fi/. See Supplementary Methods, section 1 for further details.

Some of the end points have a high number of overlapping cases. Therefore, to avoid reporting highly repetitive end points, we clustered all end points if there was an overlap of >50% of cases between them and chose the one with the most genome-wide significant hits. On a few occasions, a manual choice

was made to select the most representative end point among the correlating end points. After clustering, we had 1,932 end points for the main GWAS analysis.

Genotyping and QC

Samples were genotyped with Illumina (Illumina) and Affymetrix arrays (Thermo Fisher Scientific). Genotype calls were made with GenCall and zCall algorithms for Illumina and the AxiomGT1 algorithm for Affymetrix data. Chip genotyping data produced with previous chip platforms and reference genome builds were lifted over to build v.38 (GRCh38/hg38) following a previously described protocol[52]. In sample-wise QC, individuals with genetically inferred sex not matching the reported sex in registries, high genotype missingness (>5%) and excess heterozygosity (±4 standard deviations) were removed. In variant-wise QC, variants with high missingness (>2%), low Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$) and minor allele count < 3 were removed. Chip-genotyped samples were pre-phased with Eagle v.2.3.5 (https://data.broadinstitute.org/alkesgroup/Eagle/) using default parameters, except the number of conditioning haplotypes was set to 20,000.

Genotype imputation with a population-specific reference panel

The population-specific Sequencing Initiative Suomi (SISu) v.3 imputation reference panel was developed by using high-coverage (25–30 times) whole-genome sequencing data for 3,775 Finnish individuals. In brief, the variant call set was produced using the GATK HaplotypeCaller algorithm by following GATK best practices for variant calling. Genotype-wise, sample-wise and variant-wise QC was performed using the Hail framework (https://github.com/hail-is/hail) v.0.1, and the resulting high-quality whole-genome sequencing data were phased (Supplementary Methods). Genotype imputation was carried out using the SISu v.3 reference panel with Beagle 4.1 (v.08Jun17.d8b, https://faculty.washington.edu/browning/beagle/b4_1.html) as described in a previous protocol[53]. Post-imputation QC involved non-reference concordance analyses, checking expected conformity of the imputation INFO values distribution, MAF differences between the target dataset and

the imputation reference panel, and checking chromosomal continuity of the imputed genotype calls. After these steps, variants with imputation INFO scores of <0.6 or MAF values of <0.0001 were excluded.

Association analysis and fine-mapping

The mixed-model logistic regression method SAIGE (v.0.35.8.8)[54] was used for association analysis. We used sex, age, genotyping batch and ten PCs as covariates (see Supplementary Methods for details). We used SuSiE[55] for fine-mapping. We fine-mapped all regions with variants that had values of $P < 1 \times 10^{-6}$ and extended regions 1.5 Mb upstream and downstream from each lead variant. Finally, overlapping regions were merged and subjected to fine-mapping. The major histocompatibility complex region (chromosome 6: 25–36 Mb) was excluded owing to its complex LD structure. We allowed up to ten independent signals per region, and SuSiE reports a 95% credible set for each independent signal. As LD, we used in-sample dosages (that is, cases and controls used for each phenotype) computed with LDStore2. The FinnGen fine-mapping pipeline is available in GitHub (https://github.com/FINNGEN/finemapping-pipeline).

To define independent signals within a locus, we utilized fine-mapping results. For each locus, we report the credible set as an independent hit if it represents a primary strongest signal with lead $P < 5 \times 10^{-8}$. For secondary hits, we required genome-wide significance and log Bayes factor (BF) > 2. The BF filtering was necessary because SuSiE sometimes reports multiple credible sets for a single strong signal but this is indicated in SuSiE as a low BF (the model does not improve by adding another signal in the region that is not an independent signal).

Browser development

The https://r5.finngen.fi browser was developed based on the PheWeb[56] codebase.

Estimation of expected number of enriched variant associations

We aimed to estimate whether we observed variant associations that were enriched by more than twofold in the Finnish population in the lower frequency range (NFSEE MAF < 5%) than would be expected by chance. To this end, we sampled a subset of variants (NFSEE MAF < 5%) that were not associated with any end point in FinnGen ($P > 0.001$). We drew 1 million random samples of the number of independent hits (143) observed in a GWAS from the set of non-associated variants. To closely follow the observed frequency distribution, we further matched the random samples to contain the same number of variants in each frequency bin ((0,0.001], (0.001,0.005], (0.005,0.01] and then in 0.01 bins up to 0.05). We computed the mean and standard deviations of per cent twofold enriched variants from the random samples and calculated *P* values from the normal distribution using the randomized mean and standard deviation.

EstBB and UKBB replication

The EstBB is a population-based biobank at the Institute of Genomics, University of Tartu. The current cohort size is 200,000 individuals (aged ≥18 years), reflecting the age, sex and geographical distribution of the adult Estonian population. Overall, 83% of the samples are from Estonian individuals, 14% from Russian people and 3% from other ethnicities. All participants were recruited by general practitioners, physicians in hospitals and during promotional events. After recruitment, all participants completed a questionnaire about their health status, lifestyle and diet. Specifically, the questionnaire included personal data (place of birth, place(s) of living, nationality, among others), genealogical data (family history of medical conditions spanning four generations), educational and occupational history, and lifestyle data (physical activity, dietary habits (food frequency questionnaires), smoking status, alcohol consumption, women's health and quality of life). The EstBB database is linked with national registries (such as the Cancer Registry and Causes of Death Registry), hospital databases and the database of the national health insurance fund, which holds treatment and procedure service bills. Diseases and health problems are recorded as ICD-10 codes and prescribed medicine according to the ATC classification. These health data are continuously updated through periodical linking to national electronic databases and registries. All participants were genotyped with genome-wide chip arrays and further imputed with a

population-specific imputation panel consisting of 2,244 high-coverage (30 times) whole-genome sequence data from individuals and 16,271,975 high-quality variants[57]. Researchers at the EstBB ran an association analysis of the 15 phenotypes (Supplementary Table 8) used in this study in 136,724 individuals. The association analysis was conducted with SAIGE52 mixed models with age, sex and ten PCs used as covariates.

We used the Pan UKBB (https://pan.ukbb.broadinstitute.org/) project European subset association analysis summary statistics in the UKBB replication[58] (Supplementary Table 7).

As both the EstBB and the UKBB are on human genome build 37, we lifted over the coordinates to build 38 to match FinnGen. Variants were then matched on the basis of chromosome, position, reference and alternative alleles.

Inverse variance weighted meta-analysis was used to perform a meta-analysis on the three cohorts (code available at https://github.com/FINNGEN/META_ANALYSIS).

Variant annotation

We utilized Variant Effect Predictor (https://www.ensembl.org/info/docs/tools/vep/index.html) for annotating imputation panel variants. For coding variants, we chose a single most-severe consequence and corresponding gene among canonical transcripts. We considered stop gained, frameshift variant, splice donor, splice acceptor, missense variant, start lost, stop lost, inframe insertion and inframe deletion as coding variants. We executed the variant annotation using Hail[59].

Colocalization

We applied colocalization to all fine-mapped regions. As a colocalization approach, we used the probabilistic model for integrating GWAS and eQTL data presented in eCAVIAR[60]. Given the PIP values of each phenotype in a region of interest, we calculated the colocalization posterior probability

(CLPP). In contrast to eCAVIAR, we used SuSiE[55] to estimate the posterior inclusion probabilities.

For a pair of phenotypes, we searched for an intersection of variants between their credible sets CS_k, $k = 1\ldots k$, and computed the CLPP as follows:

$${{\rm{CLPP}}}_{k}=\sum _{{\rm{i}}}{\rm{in}}\,{{\rm{CS}}}_{k}\,{\rm{p}}{1}_{{\rm{i}}}\,\times \,{\rm{p}}{2}_{{\rm{i}}},$$

where p1 and p2 are the PIP values from phenotypes 1 and 2, respectively.

We performed colocalization between FinnGen end points, the eQTL Catalogue[61] and selected 36 continuous end points and 57 biomarkers from the UKBB[10]. eQTL Catalogue and UKBB traits were processed with a functionally equivalent fine-mapping pipeline[10] to FinnGen and ref. [61], and credible sets provided by those studies were used in colocalization.

Annotating putatively new associations

For each association lead variants, we used the Open Targets[62] API platform (https://api.platform.opentargets.org/) to search whether any genome-wide significant hits ($P < 5 \times 10^{-8}$) have been reported for the variant (or tagging LD variants $r^2 > 0.2$) in the GWAS Catalog or the UKBB as harmonized by Open Targets (annotated 19 May 2022). We also searched whether the variant was reported as pathogenic or likely pathogenic in ClinVar[63] (ClinVar release date 7 May 2022).

Automatic annotation of known GWAS hits

To identify new hits from the GWAS results, we compared the fine-mapped results against genome-wide significant hits ($P < 5 \times 10^{-8}$) in the GWAS Catalog association database[64] and manually curated genome-wide significant hits from large GWASs (Table 1). We checked and reported separately matches in credible set variants and matches with any variants in LD with a lead variant (highest PIP) after fine-mapping. LD lookup variants were chosen using the following criteria: (1) they were less than 1,500 kb

away from the lead variant; (2) they had a $P < 0.01$; (3) and their LD squared Pearson's correlation with the lead variant was higher than a dynamic LD threshold based on the P value of the lead variant so that the expected P value of the linked variant would be nominally significant ($r^2 = 5$/inverse chi-squared survival function (P value)).

A variant was considered to be already associated if its chromosome and position were identical to the GWAS Catalog association and if its reference and alternative allele matched the strand-aligned and effect-aligned association alleles. Because the GWAS Catalog associations do not have complete allele information, the allele information for associations was retrieved from dbSNP data, human genome build 153, assembly 38. The GWAS Catalog version used was released on 21 April 2021.

Ethics statement

Participants in FinnGen provided informed consent for biobank research on basis of the Finnish Biobank Act. Alternatively, separate research cohorts, collected before the Finnish Biobank Act came into effect (in September 2013) and the start of FinnGen (August 2017) were collected on the basis of study-specific consent and later transferred to the Finnish biobanks after approval by Fimea, the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) approved the FinnGen study protocol (number HUS/990/2017).

The FinnGen study is approved by the THL (approval number THL/2031/6.02.00/2017, amendments THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019 and THL/1721/5.05.00/2019), the Digital and Population Data Service Agency (VRK43431/2017-3, VRK/6909/2018-3 and VRK/4415/2019-3), the Social Insurance Institution (KELA) (KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019 and KELA 98/522/2019) and Statistics Finland (TK-53-1041-17).

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 5 include the following datasets: THL Biobank

BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8 and BB2019_26; Finnish Red Cross Blood Service Biobank 7.12.2017; Helsinki Biobank HUS/359/2017; Auria Biobank AB17-5154; Biobank Borealis of Northern Finland_2017_1013; Biobank of Eastern Finland 1186/2018; Finnish Clinical Biobank Tampere MH0004; Central Finland Biobank 1-2017; and Terveystalo Biobank STB 2018001.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Based on National and European regulations (GDPR) access to individual-level sensitive health data must be approved by national authorities for specific research projects and for specifically listed and approved researchers. The health data described here was generated and provided by the National Health Register Authorities (Finnish Institute of Health and Welfare, Statistics Finland, KELA, Digital and Population Data Services Agency) and approved, either by the individual authorities or by the Finnish Data Authority, Findata, for use in the FinnGen project. Therefore, we, the authors of this paper, are not in a position to grant access to individual-level data to others. However, any researcher can apply for the health register data from the Finnish Data Authority Findata (https://findata.fi/en/permits/) and for individual-level genotype data from Finnish biobanks via the Fingenious portal (https://site.fingenious.fi/en/) hosted by the Finnish Biobank Cooperative FINBB (https://finbb.fi/en/). All Finnish biobanks can provide access for research projects within the scope regulated by the Finnish Biobank Act, which is research utilizing the biobank samples or data for the purposes of promoting health, understanding the mechanisms of disease or developing products and treatment practices used in health and medical care. The genotype data for the FinnGen release 5 used in this study was returned to the biobanks at the same time as the public release of the FinnGen release 5 summary results was done. All summary statistics described in this manuscript can be found in the Supplementary Information. All information

regarding data download of summary statistics of additive GWAS of FinnGen release 5 can be found through the following link: https://finngen.gitbook.io/documentation/v/r5/data-download. You can learn more about accessing other FinnGen data here: https://www.finngen.fi/en/access_results. A full list of FinnGen end points for release 5 is available at: https://www.finngen.fi/en/researchers/clinical-endpoints. A full list of gene variants captured by the FinnGen specific Axiom array can be found at: https://www.finngen.fi/en/researchers/genotyping and https://www.dropbox.com/s/n8srnyy547resrq/finngen2_proposal_5_5_2019.tsv?dl=0.

Code availability

Central data analysis and processing pipelines used are freely available: fine-mapping pipeline (https://github.com/FINNGEN/finemapping-pipeline); meta-analysis (https://github.com/FINNGEN/META_ANALYSIS); genetic ancestry and PCA pipeline (https://github.com/FINNGEN/pca_kinship); and GWAS SAIGE pipeline (https://github.com/FINNGEN/saige-pipelines). Please see https://finngen.gitbook.io/documentation/ for a detailed description of data production and analysis including code used to run analyses. Please see https://github.com/FINNGEN/ for further code repositories used to run analyses in FinnGen.

R v4.0.3 (https://www.r-project.org/) was used to create plots and analyse data. R codes used to reproduce figures are available upon request.

References

1. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).

2. Xue, Y. et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat. Commun.* **8**, 15927 (2017).

3. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1322563111 (2014).

4. Norio, R. The Finnish Disease Heritage III: the individual diseases. *Hum. Genet.* **112**, 470–526 (2003).

5. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

6. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

7. Kerminen, S. et al. Fine-scale genetic structure in Finland. *G3* **7**, 3459–3468 (2017).

8. Ritari, J., Koskela, S., Hyvärinen, K., FinnGen & Partanen, J. HLA-disease association and pleiotropy landscape in over 235,000 Finns. *Hum. Immunol.* **83**, 391–398 (2022).

9. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

10. Kanai, M. et al. Insights from complex trait fine-mapping across diverse populations. Preprint at *medRxiv* https://doi.org/10.1101/2021.09.03.21262975 (2021).

11. Sakaue, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).

12. Sun, B. B. et al. Genetic associations of protein-coding variants in human disease. *Nature* **603**, 95–102 (2022).

13. Heyne, H. O. et al. Mono- and biallelic effects of on disease at biobank scale. *Nature* https://doi.org/10.1038/s41586-022-05420-7 (2022).

14. Patel, K. A. et al. Heterozygous RFX6 protein truncating variants are associated with MODY with reduced penetrance. *Nat. Commun.* **8**, 888 (2017).

15. Norberg, A. et al. Novel variants in Nordic patients referred for genetic testing of telomere-related disorders. *Eur. J. Hum. Genet.* **26**, 858–867 (2018).

16. Löf, C. et al. Detection of novel gene variants associated with congenital hypothyroidism in a Finnish patient cohort. *Thyroid* **26**, 1215–1224 (2016).

17. Jiang, X.-C. & Yu, Y. The role of phospholipid transfer protein in the development of atherosclerosis. *Curr. Atheroscler. Rep.* **23**, 9 (2021).

18. Teslovich, T. M. et al. Biological, clinical, and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

19. van der Harst, P. & Verweij, N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* **122**, 433–443 (2018).

20. Shadrina, A. S. et al. Prioritization of causal genes for coronary artery disease based on cumulative evidence from experimental and in silico studies. *Sci. Rep.* **10**, 10486 (2020).

21. Dimitra, A. et al. *HHIPL1*, a gene at the 14q32 coronary artery disease locus, positively regulates hedgehog signaling and promotes atherosclerosis. *Circulation* **140**, 500–513 (2019).

22. Bindesbøll, C. et al. NBEAL1 controls SREBP2 processing and cholesterol metabolism and is a susceptibility locus for coronary artery disease. *Sci. Rep.* **10**, 4528 (2020).

23. Graydon, C. G., Mohideen, S. & Fowke, K. R. LAG3's enigmatic mechanism of action. *Front. Immunol.* https://doi.org/10.3389/fimmu.2020.615317 (2021).

24. FDA approves anti-LAG3 checkpoint. *Nat. Biotechnol.* **40**, 625 (2022).

25. Boer, C. G. et al. Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations. *Cell* **184**, 4784–4818.e17 (2021).

26. Sliz, E. et al. Uniting biobank resources reveals novel genetic pathways modulating susceptibility for atopic dermatitis. *J. Allergy Clin. Immunol.* **149**, 1105–1112.e9 (2022).

27. Li, J. et al. Utility of basophil activation test for predicting the outcome of wheezing in children: a pilot study. *BMC Immunol.* **22**, 4 (2021).

28. Chakraborti, S., Natarajan, K., Curiel, J., Janke, C. & Liu, J. The emerging role of the tubulin code: from the tubulin molecule to neuronal function and disease. *Cytoskeleton* **73**, 521–550 (2016).

29. Sharifinejad, N. et al. Clinical, immunological, and genetic features in 49 patients with ZAP-70 deficiency: a systematic review. *Front. Immunol.* **11**, 831 (2020).

30. Mulford, A. J., Wing, C., Dolan, M. E. & Wheeler, H. E. Genetically regulated expression underlies cellular sensitivity to chemotherapy in diverse populations. *Hum. Mol. Genet.* **30**, 305–317 (2021).

31. Mafra, F. et al. Copy number variation analysis reveals additional variants contributing to endometriosis development. *J. Assist. Reprod. Genet.* **34**, 117–124 (2017).

32. Koskela, J. T. et al. Genetic variant in *SPDL1* reveals novel mechanism linking pulmonary fibrosis risk and cancer protection. Preprint at *medRxiv* https://doi.org/10.1101/2021.05.07.21255988 (2021).

33. Ruotsalainen, S. E. et al. Inframe insertion and splice site variants in *MFGE8* associate with protection against coronary atherosclerosis. *Commun. Biol.* **5**, 802 (2022).

34. Rämö, J. T. et al. Genome-wide screen of otosclerosis in population biobanks: 27 loci and shared associations with skeletal structure *Nat. Commun.* https://doi.org/10.1038/s41467-022-32936-3 (2023).

35. Tanigawa, Y. et al. Rare protein-altering variants in *ANGPTL7* lower intraocular pressure and protect against glaucoma. *PLoS Genet.* **16**, e1008682 (2020).

36. Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).

37. Ehrhardt, N. et al. Hepatic Tm6sf2 overexpression affects cellular ApoB-trafficking, plasma lipid levels, hepatic steatosis and atherosclerosis. *Hum. Mol. Genet.* **26**, 2719–2731 (2017).

38. Prill, S. et al. The *TM6SF2* E167K genetic variant induces lipid biosynthesis and reduces apolipoprotein B secretion in human hepatic 3D spheroids. *Sci. Rep.* **9**, 11585 (2019).

39. Pirola, C. J. & Sookoian, S. The dual and opposite role of the *TM6SF2*-rs58542926 variant in protecting against cardiovascular disease and conferring risk for nonalcoholic fatty liver: a meta-analysis. *Hepatology* **62**, 1742–1756 (2015).

40. Buch, S. et al. A genome-wide association study confirms *PNPLA3* and identifies *TM6SF2* and *MBOAT7* as risk loci for alcohol-related cirrhosis. *Nat. Genet.* **47**, 1443–1448 (2015).

41. Tang, S. et al. Association of *TM6SF2* rs58542926 T/C gene polymorphism with hepatocellular carcinoma: a meta-analysis. *BMC Cancer* **19**, 1128 (2019).

42. Kim, D. S. et al. Novel association of *TM6SF2* rs58542926 genotype with increased serum tyrosine levels and decreased apoB-100 particles in Finns. *J. Lipid Res.* **58**, 1471–1481 (2017).

43. Mars, N. et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* **26**, 549–557 (2020).

44. Kiiskinen, T. et al. Genomic prediction of alcohol-related morbidity and mortality. *Transl Psychiatry* **10**, 23 (2020).

45. Strausz, S. et al. Genetic analysis of obstructive sleep apnoea discovers a strong association with cardiometabolic health. *Eur. Respir. J.* **57**, 2003091 (2021).

46. Helkkula, P. et al. ANGPTL8 protein-truncating variant associated with lower serum triglycerides and risk of coronary disease. *PLoS Genet.* **17**, e1009501 (2021).

47. Rahimov, F. et al. High incidence and regional distribution of cleft palate in Finns are associated with a functional variant in an *IRF6* enhancer. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-941741/v1 (2021).

48. Niemi, M. E. K. et al. Mapping the human genetic architecture of COVID-19. *Nature* https://doi.org/10.1038/s41586-021-03767-x (2021).

49. Zhou, W. et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human disease. *Cell Genom.* **2**, 100192 (2022).

50. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* **10**, 190221 (2020).

51. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

52. Pärn, K. et al. Genotyping chip data lift-over to reference genome build GRCh38/hg38 V.2. *protocols.io* https://doi.org/10.17504/protocols.io.nqtddwn (2019).

53. Palta, P. Genotype imputation workflow v3.0 V.1. *protocols.io* https://doi.org/10.17504/protocols.io.nmndc5e (2018).

54. Zhou, W. et al. Efficiently controlling for case–control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

55. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic

fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

56. Gagliano Taliun, S. A. Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).

57. Mitt, M. et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).

58. *Pan-UK Biobank* (Pan UK Biobank Team, 2020); https://pan.ukbb.broadinstitute.org.

59. Hail v.0.2 (Hail Team, 2019); https://github.com/hail-is/hail.

60. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

61. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).

62. Ochoa, D. et al. Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).

63. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

64. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

Acknowledgements

We thank all those who contributed samples and data for the FinnGen scientific project; and P. VandeHaar for technical consultation on PheWeb. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie, AstraZeneca UK, Biogen, Bristol Myers Squibb (and Celgene Corporation & Celgene International II), Genentech, Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA, Pfizer, GlaxoSmithKline Intellectual Property Development, Sanofi US Services, Maze Therapeutics, Janssen Biotech, Novartis, and Boehringer Ingelheim. The following biobanks are acknowledged for delivering samples to FinnGen: Auria Biobank (https://www.auria.fi/biopankki/), THL Biobank (https://www.thl.fi/biobank), Helsinki Biobank (https://www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (https://www.ita-suomenbiopankki.fi/en), Central Finland Biobank (https://www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank (https://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish biobanks are members of the BBMRI.fi infrastructure (https://www.bbmri.fi). The FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious services (https://site.fingenious.fi/en/) managed by FINBB. We acknowledge the BBMRI.fi for creating the original biobank map used in Fig. 1a and providing us their permission to use it.

Funding

Open Access funding provided by University of Helsinki including Helsinki University Central Hospital.

Author information

Author notes

1. These authors jointly supervised this work: Mark J. Daly, Aarno Palotie

Authors and Affiliations

1. Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland

   Mitja I. Kurki, Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati M. Donner, Mary P. Reeve, Hannele Laivuori, Mervi Aavikko, Mari A. Kaunisto, Elisa Lahtela, Pietro Della Briotta Parolo, Arto A. Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O. Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, Sanni Ruotsalainen, Kalle Pärn, Vincent Llorens, Oluwaseun A. Dada, Andrea Ganna, Awaisa Ghazal, Risto Kajanne, Jaakko Kaprio, Elina Kilpeläinen, Aoxing Liu, Mari E. K. Niemi, Hanna M. Ollila, Rodosthenis Rodosthenous, Elmo Saarentaus, Huei-yi Shen, Taru Tukiainen, Tiinamaija Tuomi, Aki S. Havulinna, Samuli Ripatti, Anu Jalanko, Mark J. Daly & Aarno Palotie

2. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

   Mitja I. Kurki, Juha Karjalainen, Masahiro Kanai, Henrike O. Heyne, Kumar Veerapen, Wei Zhou, Hilary Finucane, Andrea Ganna, Jacob C. Ulirsch, Mark J. Daly & Aarno Palotie

3. Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA

Mitja I. Kurki, Juha Karjalainen, Masahiro Kanai, Henrike O. Heyne, Kumar Veerapen, Wei Zhou, Hilary Finucane, Andrea Ganna, Jacob C. Ulirsch, Mark J. Daly & Aarno Palotie

4. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

Mitja I. Kurki, Juha Karjalainen, Masahiro Kanai, Kumar Veerapen, Wei Zhou, Hilary Finucane, Andrea Ganna, Mark J. Daly & Aarno Palotie

5. Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia

Priit Palta, Kadri Reis, Abdelrahman G. Elnahas, Triin Laisk, Reedik Mägi, Anu Reigo, Andres Metspalu, Lili Milani & Tõnu Esko

6. Finnish Institute for Health and Welfare (THL), Helsinki, Finland

Kati Kristiansson, Hannele Mattsson, Päivi Laiho, Susanna Lemmelä, Tero Hiekkalinna, Sami Koskelainen, Teemu Paajanen, Johannes Kettunen, Terhi Kilpi, Teemu Niiranen, Veikko Salomaa, Kaisa Silander, Sirpa Soini, Felix Vaura, Aki S. Havulinna & Markus Perola

7. Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

Hannele Laivuori

8. Department of Obstetrics and Gynecology, Tampere University Hospital, Tampere, Finland

Hannele Laivuori

9. Faculty of Medicine and Health Technology, Center for Child, Adolescent and Maternal Health, University of Tampere, Tampere, Finland

Hannele Laivuori

10. Helsinki Biobank, University of Helsinki and Hospital District of Helsinki and Uusimaa, Helsinki, Finland

Anu Loukola, Olli Carpen & Eero Punkka

11. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Masahiro Kanai

12. Digital Health Center, Hasso Plattner Institute for Digital Engineering, University of Potsdam Potsdam, Potsdam, Germany

Henrike O. Heyne

13. Hasso Plattner Institute for Digital Health at Mount Sinai, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Henrike O. Heyne

14. TAUCHI Research Center, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

Javier Gracia-Tabuenca, Harri Siirtola & Marianna Niemi

15. Translational Biology, Research and Development, Biogen, Cambridge, MA, USA

Benjamin Sun

16. BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

Benjamin Sun

17. Optima Partners, Edinburgh, UK

Christopher N. Foley

18. MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK

Christopher N. Foley

19. Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

Katriina Aalto-Setälä & Mika Kähönen

20. Institute of Computer Science, University of Tartu, Tartu, Estonia

Kaur Alasoo & Nurlan Kerimov

21. Finnish Red Cross Blood Service, Helsinki, Finland

Mikko Arvas & Jukka Partanen

22. GlaxoSmithKline, Espoo, Finland

Kirsi Auro

23. Bristol Myers Squibb, New York, NY, USA

Shameek Biswas, Joseph Maranville & Robert Plenge

24. Tampere University Hospital and Tampere University, Tampere, Finland

Argyro Bizaki-Vallaskangas

25. Biogen, Cambridge, MA, USA

Chia-Yen Chen, Sally John & Heiko Runz

26. Boehringer Ingelheim, Ingelheim am Rhein, Germany

Zhihao Ding, Jan-Nygaard Jensen, Marc Jung, Nathan Lawless & George Okafo

27. GlaxoSmithKline, Collegeville, PA, USA

Margaret G. Ehm

28. Division of Rheumatology, Department of Medicine, Helsinki University Central Hospital, Helsinki, Finland

Kari Eklund

29. Orton Orthopedic Hospital, Helsinki, Finland

Kari Eklund

30. Abdominal Center, Helsinki University Hospital, Helsinki University, Helsinki, Finland

Martti Färkkilä

31. Maze Therapeutics, South San Francisco, CA, USA

Robert R. Graham, Eric M. Green & Sahar V. Mozaffari

32. Auria Biobank, University of Turku and Turku University Hospital, Turku, Finland

Antti Hakanen, Lila Kallio, Johanna Schleutker & Petri Virolainen

33. FINBB, Finnish Biobank Cooperative, Helsinki, Finland

Marco Hautalahti

34. Pfizer, New York, NY, USA

Åsa K. Hedman & Xinli Hu

35. Department of Medicine, Karolinska Institute, Solna, Sweden

Åsa K. Hedman

36. Clinical Biobank Tampere, Tampere University and Tampere University Hospital, Tampere, Finland

Mikko Hiltunen

37. Medical Research Center Oulu and PEDEGO Research Unit, University of Oulu, Oulu, Finland

Reetta Hinttala

38. Biocenter Oulu, University of Oulu, Oulu, Finland

Reetta Hinttala & Johannes Kettunen

39. Oulu University Hospital, Oulu, Finland

Reetta Hinttala

40. Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Iiris Hovatta

41. SleepWell Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Iiris Hovatta

42. Merck & Co, Kenilworth, NJ, USA

Adriana Huertas-Vazquez, Simonne Longerich, Neha Raghavan, Jae Hoon Sul & Caroline Fox

43. PEDEGO Research Unit, University of Oulu, Oulu, Finland

Laura Huilaja & Kaisa Tasanen

44. Department of Dermatology and Medical Research Center Oulu, Oulu University Hospital, Oulu, Finland

Laura Huilaja & Kaisa Tasanen

45. Genentech, San Francisco, CA, USA

Julie Hunkapiller, Rion K. Pendergrass & Mark McCarthy

46. AbbVie, Chicago, IL, USA

Howard Jacob, Fedik Rahimov, Bridget Riley-Gillis & Jeffrey Waring

47. Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Heikki Joensuu, Tuomo Meretoja & Joni A. Turunen

48. Neuro Center, Neurology, Kuopio University Hospital, Kuopio, Finland

Valtteri Julkunen

49. Institute of Clinical Medicine–Neurology, University of Eastern Finland, Kuopio, Finland

Valtteri Julkunen

50. Northern Finland Biobank Borealis, University of Oulu, Northern Ostrobothnia Hospital District, Oulu, Finland

Juhani Junttila & Raisa Serpi

51. Department of Ophthalmology, Kuopio University Hospital, Kuopio, Finland

Kai Kaarniranta

52. Department of Ophthalmology, Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland

Kai Kaarniranta

53. Department of Clinical Physiology, Tampere University Hospital, Tampere, Finland

Mika Kähönen

54. Epilepsy Center, Kuopio University Hospital, Kuopio, Finland

Reetta Kälviäinen

55. Department of Neurology, University of Eastern Finland, Kuopio, Finland

Reetta Kälviäinen

56. Department of Public Health, University of Helsinki, Helsinki, Finland

Jaakko Kaprio

57. Computational Medicine, Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland

Johannes Kettunen

58. Translational Sciences, Sanofi R&D, Framingham, MA, USA

Katherine Klinger & Deepak Rajpal

59. Biobank of Eastern Finland, University of Eastern Finland, Kuopio, Finland

Veli-Matti Kosma & Arto Mannermaa

60. Kuopio University Hospital, Kuopio, Finland

Veli-Matti Kosma & Arto Mannermaa

61. Central Finland Biobank, Central Finland Health Care District, Jyväskylä, Finland

Teijo Kuopio & Jari Laukkanen

62. Department of Clinical Genetics, Tampere University Hospital, Tampere, Finland

Venla Kurra

63. Department of Clinical Genetics, Faculty of Medicine and Health Technology, Tampere, Finland

Venla Kurra

64. Department of Medicine, Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland

Jari Laukkanen

65. FINBB, Finnish Biobank Cooperative, Turku, Finland

Johanna Mäkelä

66. Department of Otorhinolaryngology–Head and Neck Surgery, University of Helsinki, Helsinki, Finland

Antti Mäkitie

67. Helsinki University Hospital, Helsinki, Finland

Antti Mäkitie & Eveliina Salminen

68. Pfizer, Cambridge, MA, USA

Anders Malarstig

69. Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Solna, Sweden

Anders Malarstig

70. Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK

Athena Matakidou, Dirk S. Paul, Slavé Petrovski & Ioanna Tachmazidou

71. Turku University Hospital and University of Turku, Turku, Finland

Teemu Niiranen & Antti Palomäki

72. Novartis Institutes for BioMedical Research, Cambridge, MA, USA

Christopher J. O´Donnell, Ma´en Obeidat & Nicole A. Renaud

73. Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA

Hanna M. Ollila

74. Department of Oral and Maxillofacial Diseases, Helsinki University Hospital, Helsinki, Finland

Tuula Palotie

75. Department of Oral and Maxillofacial Diseases, University of Helsinki, Helsinki, Finland

Tuula Palotie, Pirkko Pussinen & Aino Salminen

76. Finnish Hematological Biobank, Helsinki, Finland

Jukka Partanen

77. Department of Pulmonary Diseases, Kuopio University Hospital, Kuopio, Finland

Margit Pelkonen

78. Department of Otorhinolaryngology, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Anne Pitkäranta

79. Translational Science and Experimental Medicine, Research and Early Development, Respiratory and Immunology, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK

Adam Platt

80. GlaxoSmithKline, Stevenage, UK

David Pulford

81. Department of Clinical Genetics, HUSLAB, HUS Diagnostic Center, University of Helsinki, Helsinki, Finland

Eveliina Salminen

82. Novartis Institutes for BioMedical Research, Basel, Switzerland

Richard Siegel

83. Finnish Clinical Biobank Tampere, Tampere University and Tampere University Hospital, Tampere, Finland

Sanna Siltanen & Tarja Laitinen

84. Department of Neurology, Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland

Hilkka Soininen

85. Department of Neurology, Helsinki University Hospital, Helsinki, Finland

Pentti Tienari

86. Translational Immunology, Research Programs Unit, University of Helsinki, Helsinki, Finland

Pentti Tienari

87. Department of Allergy, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Sanna Toppila-Salmi

88. Abdominal Center, Endocrinology, Helsinki University Hospital, Helsinki, Finland

Tiinamaija Tuomi

89. Folkhalsan Research Center, Helsinki, Finland

Tiinamaija Tuomi

90. Research Program of Clinical and Molecular Metabolism, University of Helsinki, Helsinki, Finland

Tiinamaija Tuomi

91. Eye Genetics Group, Folkhälsan Research Center, Helsinki, Finland

Joni A. Turunen

92. University of Turku, Turku, Finland

Felix Vaura

93. Janssen Research & Development, Spring House, PA, USA

Dawn Waterworth

94. Janssen Biotech, Beerse, Belgium

Robert Yang

95. Genomics Core Facility, Institute of Genomics, University of Tartu, Tartu, Estonia

Mari Nelis

96. Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland

Tomi P. Mäkelä

Consortia

FinnGen

- Mitja I. Kurki
- , Juha Karjalainen
- , Priit Palta
- , Timo P. Sipilä
- , Kati Kristiansson
- , Kati M. Donner
- , Mary P. Reeve
- , Hannele Laivuori
- , Mervi Aavikko
- , Mari A. Kaunisto
- , Anu Loukola
- , Elisa Lahtela
- , Hannele Mattsson
- , Päivi Laiho
- , Pietro Della Briotta Parolo
- , Arto A. Lehisto
- , Masahiro Kanai
- , Nina Mars

- , Joel Rämö
- , Tuomo Kiiskinen
- , Henrike O. Heyne
- , Kumar Veerapen
- , Sina Rüeger
- , Susanna Lemmelä
- , Wei Zhou
- , Sanni Ruotsalainen
- , Kalle Pärn
- , Tero Hiekkalinna
- , Sami Koskelainen
- , Teemu Paajanen
- , Vincent Llorens
- , Javier Gracia-Tabuenca
- , Harri Siirtola
- , Benjamin Sun
- , Katriina Aalto-Setälä
- , Mikko Arvas
- , Kirsi Auro
- , Shameek Biswas
- , Argyro Bizaki-Vallaskangas
- , Olli Carpen
- , Chia-Yen Chen
- , Oluwaseun A. Dada
- , Zhihao Ding
- , Margaret G. Ehm
- , Kari Eklund
- , Martti Färkkilä
- , Hilary Finucane
- , Andrea Ganna
- , Awaisa Ghazal
- , Robert R. Graham
- , Eric M. Green
- , Antti Hakanen
- , Marco Hautalahti
- , Åsa K. Hedman
- , Mikko Hiltunen

- , Reetta Hinttala
- , Iiris Hovatta
- , Xinli Hu
- , Adriana Huertas-Vazquez
- , Laura Huilaja
- , Julie Hunkapiller
- , Howard Jacob
- , Jan-Nygaard Jensen
- , Heikki Joensuu
- , Sally John
- , Valtteri Julkunen
- , Marc Jung
- , Juhani Junttila
- , Kai Kaarniranta
- , Mika Kähönen
- , Risto Kajanne
- , Lila Kallio
- , Reetta Kälviäinen
- , Jaakko Kaprio
- , Nurlan Kerimov
- , Johannes Kettunen
- , Elina Kilpeläinen
- , Terhi Kilpi
- , Katherine Klinger
- , Veli-Matti Kosma
- , Teijo Kuopio
- , Venla Kurra
- , Jari Laukkanen
- , Nathan Lawless
- , Aoxing Liu
- , Simonne Longerich
- , Johanna Mäkelä
- , Antti Mäkitie
- , Anders Malarstig
- , Arto Mannermaa
- , Joseph Maranville
- , Athena Matakidou

- , Tuomo Meretoja
- , Sahar V. Mozaffari
- , Mari E. K. Niemi
- , Marianna Niemi
- , Teemu Niiranen
- , Christopher J. O´Donnell
- , Ma´en Obeidat
- , George Okafo
- , Hanna M. Ollila
- , Antti Palomäki
- , Tuula Palotie
- , Jukka Partanen
- , Dirk S. Paul
- , Margit Pelkonen
- , Rion K. Pendergrass
- , Slavé Petrovski
- , Anne Pitkäranta
- , Adam Platt
- , David Pulford
- , Eero Punkka
- , Pirkko Pussinen
- , Neha Raghavan
- , Fedik Rahimov
- , Deepak Rajpal
- , Nicole A. Renaud
- , Bridget Riley-Gillis
- , Rodosthenis Rodosthenous
- , Elmo Saarentaus
- , Aino Salminen
- , Eveliina Salminen
- , Veikko Salomaa
- , Johanna Schleutker
- , Raisa Serpi
- , Huei-yi Shen
- , Richard Siegel
- , Kaisa Silander
- , Sanna Siltanen

- , Sirpa Soini
- , Hilkka Soininen
- , Jae Hoon Sul
- , Ioanna Tachmazidou
- , Kaisa Tasanen
- , Pentti Tienari
- , Sanna Toppila-Salmi
- , Taru Tukiainen
- , Tiinamaija Tuomi
- , Joni A. Turunen
- , Jacob C. Ulirsch
- , Felix Vaura
- , Petri Virolainen
- , Jeffrey Waring
- , Dawn Waterworth
- , Robert Yang
- , Caroline Fox
- , Aki S. Havulinna
- , Markus Perola
- , Samuli Ripatti
- , Anu Jalanko
- , Tarja Laitinen
- , Tomi P. Mäkelä
- , Robert Plenge
- , Mark McCarthy
- , Heiko Runz
- , Mark J. Daly
- & Aarno Palotie

Contributions

M.I.K.: FinnGen analysis team leader, data analysis, browser development and wrote the first draft of the manuscript. J. Karjalainen: analysis, browser development and paper editing. P. Palta: paper writing and FinnGen genotype data development. T.P.S.: team leader. K. Kristiansson: paper editing, FinnGen phenotype and register data development. K.M.D.: paper editing and data QC. M.P.R.: paper editing and visualizations. H.L.: paper

editing, study design and FinnGen phenotype and register data development. M. Aavikko: paper editing, study design and team coordination. M.A.K.: paper editing, visualizations and variant novelty annotation. A. Loukola: paper editing, methods, biobank samples and Supplementary Methods, section 2. E.L.: paper editing, FinnGen phenotype and register data development and clinical team coordination. H.M.: FinnGen phenotype and register data development. P.L.: overseeing and organizing sample preparation and logistics. P.D.B.P.: computation tool development and population structure analysis. A.A.L.: computation tool development and fine-mapping. M. Kanai: provided fine-mapped GTEx v.8 data, UKBB quantitative phenotype fine-mapping and contributed the original fine-mapping pipeline. N.M.: paper editing. J.R.: paper editing, and ear, nose and throat end point interpretation. T. Kiiskinen: paper editing, FinnGen phenotype and register data development and data analysis. H.O.H.: paper editing and annotating and interpreting ClinVar variants. K.V.: paper editing and finalization, and initial phenotype summary statistics curation. S. Rüeger: tool development and colocalization analysis. S. Lemmelä: paper editing, and creating all phenotype plots and tables for the paper. W.Z.: GWAS analysis contribution. S. Ruotsalainen: association analysis. K.P.: genotype data development. T.H.: end point development and register data processing. S.K.: register data processing. T. Paajanen: data processing. V.L.: end point development, tool development and construction of RISTEYS. J.G.-T.: paper editing and longitudinal data analysis. H. Siirtola: paper editing and visualization of registry data in FinnGen end points. K.R.: replication analysis and paper editing. A.G.E.: replication analysis. B.S.: simulation analyses for matched sample size comparison to the UKBB and the EstBB data and paper editing. C.N.F.: simulation analyses for matched sample size comparison to the UKBB and the EstBB data and paper editing. K.A.-S.: paper editing. K. Alasoo: paper editing and fine-mapping of the eQTL Catalogue data. M. Arvas: phenotype and HLA analysis and paper editing. K. Auro: study design, overseeing the project and paper editing. S.B.: paper editing. A.B.-V.: paper editing and end point development. O.C.: paper editing, study design and end point development. C.-Y.C.: data analysis and paper editing. O.A.D.: data processing and cloud administration. Z.D.: study design, data analysis and paper editing. M.G.E.: study design, overseeing the project and paper editing. K.E.: end point development and paper editing. M.F.: end point

development, study design and paper editing. H.F.: provided fine-mapped GTEx v.8 data and UKBB quantitative phenotypes and paper editing. A. Ganna: study design, overseeing tool development and paper editing. A. Ghazal: participated in R4 imputation, full responsibility of running R5 production runs and paper editing. R.R.G.: study design, overseeing the project and paper editing. E.M.G.: study design, overseeing the project and paper editing. A.H.: study design, overseeing the project and paper editing. M. Hautalahti: biobank coordination. Å.K.H.: data analysis and paper editing. M. Hiltunen: cohort collection, clinical expertise and paper editing. R. Hinttala: data analysis and paper editing. I.H.: end point development and paper editing. X.H.: data analysis and paper editing. A.H.-V.: data analysis and paper editing. L.H.: end point development and paper editing. J.H.: study design, overseeing the project and paper editing. H. Jacob: study design, overseeing the project and paper editing. J.-N.J.: study design and paper editing. H.J.: end point development and paper editing. S.J.: study design, overseeing the project and paper editing. V.J.: end point development, study design and paper editing. M.J.: data analysis and paper editing. J.J.: study design, overseeing the project and paper editing. K. Kaarniranta: end point develpoment and paper editing. M. Kähönen: end point development, overseeing the project and paper editing. R. Kajanne: project coordination and paper editing. L.K.: sample collection and coordination. R. Kälviäinen: end point develpoment and paper editing. J. Kaprio: study design, register expertise and paper editing. N.K.: peformed fine-mapping on the eQTL Catalogue data and paper editing. J. Kettunen: data analysis and paper editing. E.K.: imputation and QC of R1–R5 datasets and paper editing. T. Kilpi: study design, overseeing the project and paper editing. K. Klinger: study design, overseeing the project and paper editing. V.-M.K.: study design, overseeing the project and paper editing. T. Kuopio: sample collection and coordination. V.K.: end point development and paper editing. T. Laisk: data analysis and paper editing. J.L.: study design, overseeing the project and paper editing. N.L.: study design, overseeing the project and paper editing. A. Liu: data analysis and paper editing. S. Longerich: study design, overseeing the project and paper editing. R.M.: genotype data preparation and paper editing. J. Mäkelä: sample collection and coordination. A. Mäkitie: end point development and paper editing. A. Malarstig: study design, overseeing the project and paper editing. A. Mannermaa: study design, overseeing the project and paper editing. J.

Maranville: study design, overseeing the project and paper editing. A. Matakidou: study design, overseeing the project and paper editing. T.M.: end point development and paper editing. S.V.M.: data analysis and paper editing. M.E.K.N.: study design, study coordination and paper editing. M. Niemi: tool and end point development and paper editing. T.N.: end point development and paper editing. C.J.O.: study design, overseeing the project and paper editing. M.O.: data analysis and paper editing. G.O.: study design, overseeing the project and paper editing. H.M.O.: data analysis and paper editing. A.P.: end point development and validation, data analysis, study design and paper editing. T. Palotie: clinical definitions and end point development and paper editing. J.P.: study design, overseeing the project, HLA imputation and paper editing. D.S.P.: study design, overseeing the project and paper editing. M. Pelkonen: end point development and paper editing. R.K.P.: data analysis and paper editing. S.P.: study design and paper editing. A. Pitkäranta: study design, overseeing the project and paper editing. A. Platt: study design, overseeing the project and paper editing. D.P.: data analysis and paper editing. E.P.: sample collection and coordination. P. Pussinen: end point definition and paper editing. N.R.: data analysis and paper editing. F.R.: data analysis and paper editing. D.R.: data analysis and paper editing. N.A.R.: data analysis and paper editing. B.R.-G.: study design, overseeing the project and paper editing. R.R.: study design, study coordination and paper editing. E. Saarentaus: paper editing, and ear, nose and throat end point interpretation. A.S.: end point development and paper editing. E. Salminen: end point development, study design and paper editing. V.S.: end point development, study design and paper editing. J.S.: end point development and paper editing. R. Serpi: sample collection and coordination. H.-y.S.: project coordination and paper editing. R. Siegel: study design and paper editing. K.S.: biobank expertise and paper editing. S. Siltanen: study design, overseeing the project and paper editing. S. Soini: biobank and legal expertise, and paper editing. H. Soininen: end point development, clinical expertise and paper editing. J.H.S.: data analysis and paper editing. I.T.: data analysis and paper editing. K.T.: end point development, clinical expertise and paper editing. P.T.: neurology end point definitions, paper review and paper editing. S.T.-S.: paper editing. T. Tukiainen: study design, data analysis and paper editing. T. Tuomi: end point development, study design and paper editing. J.A.T.: ophthalmology end point definitions, paper review and paper editing.

J.C.U.: fine-mapping and paper editing. F.V.: end point development and paper editing. P.V.: study design, overseeing the project and paper editing. J.W.: study design, data analysis and paper editing. D.W.: study design and paper editing. R.Y.: study design and paper editing. M. Nelis: study design and paper editing. A.R.: study design and paper editing. A. Metspalu: study design, Head of the Estonian Biobank and paper editing. L.M.: study design and paper editing. T.E.: study design and paper editing. C.F.: study design, overseeing the project and paper editing. A.S.H.: paper editing, FinnGen end point concept and implementation and register data development. M. Perola: paper editing, study design, ethics expert and THL science co-ordinator. S. Ripatti: paper editing, study design and data analysis expertise. A.J.: paper editing and participation in strategic planning of the project. T. Laitinen: paper editing, study design and overseeing the project. T.P.M.: paper editing, study design and overseeing the project. R.P.: paper editing, study design and overseeing the project. M.M.: paper writing and editing, study design and overseeing the project. H.R.: paper writing and editing, study design and overseeing the project. M.J.D.: paper writing and editing, study design and overseeing the project. A. Palotie: paper writing and editing, study design and overseeing the project.

Corresponding author

Correspondence to Aarno Palotie.

Ethics declarations

Competing interests

J.C.U.: has received compensation for consulting from Goldfinch Bio and is an employee of Patch Biosciences. K.E.: Consultation fees from Sobi, and Orion corporation. A. Palomäki: consulting fee from Abbvie, Amgen and Pfizer, lecture fee from Pfizer and Sanofi. H. Jacob: employee of AbbVie. F.R.: employee of AbbVie. B.R.-G.: employee of AbbVie. J.W.: employee of AbbVie. A. Matakidou: employee of AstraZeneca. D.S.P.: employee of AstraZeneca. S.P.: employee of AstraZeneca. A. Platt: employee of AstraZeneca. I.T.: employee of AstraZeneca. B.S.: employee of Biogen. C.-

Y.C.: employee of Biogen. S.J.: employee of Biogen. H.R.: employee of Biogen. Z.D.: employee of Boehringer Ingelheim. J.-N.J.: employee of Boehringer Ingelheim. M.J.: employee of Boehringer Ingelheim. N.L.: employee of Boehringer Ingelheim. G.O.: employee of Boehringer Ingelheim. S.B.: employee of Bristol Myers Squibb. J. Maranville: employee of Bristol Myers Squibb. R.P.: employee of Bristol Myers Squibb. J.H.: employee of Genentech. R.K.P.: employee of Genentech. M.M.: employee of Genentech. K. Auro: employee of GlaxoSmithKline. M.G.E.: employee of GlaxoSmithKline. D.P.: employee of GlaxoSmithKline. R.Y.: employee of Janssen Biotech. D.W.: employee of Janssen Research & Development. R.R.G.: employee of Maze Therapeutics. E.M.G.: employee of Maze Therapeutics. S.V.M.: employee of Maze Therapeutics. A.H.-V.: employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. S. Longerich: employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. N.R.: employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. J.H.S.: employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. C.F.: employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. M.E.K.N.: employee of Novartis. C.J.O.: employee of Novartis. M.O.: employee of Novartis. N.R.: employee of Novartis. R. Siegel: employee of Novartis. Å.K.H.: employee of Pfizer. X.H.: employee of Pfizer. A. Malarstig: employee of Pfizer. K. Klinger: employee of Sanofi. D.R.: employee of Sanofi. J.P.: employee of the Finnish Red Cross Blood Service. A.G.E.: employee of University of Tartu. M.J.D.: founder of Maze Therapeutics. T. Kuopio: lecture fee from Amgen, Roche and MSD. J.A.T.: lecture fee from Blueprint Genetics Finland, and on the advisory board of Novartis Finland (unrelated to this work). P.T.: lecture or consulting fee from Alexion, Roche, Merck, Janssen-Cilag, Novartis and Biogen (unrelated to this work). M. Aavikko: lecturing fee and congress trip reimbursement from Ipsen and Novartis (unrelated to this work). A. Palotie: member of the Pfizer Genetics Scientific Advisory Panel. V.S.: received a honorarium from Sanofi for consulting, and ongoing research collaboration with Bayer (all unrelated to the present study). H. Joensuu: senior consultant for Orion Pharma Orion, Chairman of the Scientific Advisory Board in Neutron Therapeutics, received a fee for acting as the Chair of an ESMO meeting from Deciphera Pharmaceuticals,

stock ownership in Orion Pharma and Sartar Therapeutics. The other authors declare no competing interests.

Peer review

Peer review information

Nature thanks Francesco Cucca, Lea Davis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 FinnGen Age Distribution and Registers.

A) Distribution of the current age (age at the end of the follow-up) and age of death for FinnGen participants B) Follow-up time and main coding used in each register among FinnGen participants in FinnGen release 5. Abbreviations: CANCER = The Finnish Cancer Registry; DEATH = Cause of death register; INPATIENT = HILMO - Care Register for Health Care: Inpatient hospital visits; OUTPATIENT = HILMO - Care Register for Health Care: Specialty outpatient visits and day surgeries; PURCHASE = Drug Purchases: All Prescription drug purchases; REIMBURSEMENT = Drug Reimbursement: entitlements for prescription drug reimbursement for certain chronic diseases.

Extended Data Fig. 2 PCA classification of 224,737 FinnGen participants combined with 1000 genomes samples (AFR,AMR,EAST,EUR,FIN,SAS).

FinnGen outlier samples were removed as deviating from the bulk of the FinnGen samples.

Extended Data Fig. 3 Comparison of effect sizes between biobanks.

A,B) Effect size (log(OR), beta) comparison of 275 genome-wide significant lead variants identified in FinnGen among 15 analysed diseases in Estonia and UKBB. The sign of beta is aligned to be positive in Estonia and UKBB. C,D) beta comparison of variants only in known loci. E,F) beta comparison of novel loci. Dashed lines indicates identity line and solid lines are the regression line (red line and text weighted by pooled standard error of betas).

Extended Data Fig. 4 Enrichment of 493 unique phenome-wide significant associations binned by NFSEE MAF and split by whether 95% credible sets contain a coding variant.

The p-values of the test of difference in average enrichment are shown on the right side of each MAF bin. Lines indicate 95% confidence interval of the mean enrichment. Number of coding/non-coding variants in each bin : 21/27, 12/35, 11/22, 3/10, 7/24 and 19/277 given in the same order as in the figure x-axis.

Supplementary information

Supplementary Information

Comparison of effects size in known genome-wide significant loci between FinnGen and large published reference GWASs (Table 1). The *y* and *x* axes represent FinnGen and reference GWAS effect sizes, respectively. Beta values are aligned to be positive in reference studies. Lines extending from points indicate standard errors in respective studies. Regression lines omit the intercept and two types of regressions are provided: unweighted and weighted by pooled standard errors from the two studies. The solid line indicates the identity line and the dotted line and dashed lines indicate

unweighted and weighted regression, respectively. Only variants with $P < 1 \times 10^{-10}$ in the reference study were included to mitigate the effect of the winner's curse of inflated beta values in the reference studies.

Reporting Summary

Supplementary Data 1

List of variants, summary statistics and references for figures in the Supplementary Information file.

Supplementary Tables 1–12

Supplementary Tables 1–12 and table legends.

Supplementary Methods

The file contains the following sections: phenotyping from nationwide population-based health registers; FinnGen participant recruitment and legacy cohorts, genotyping and genotype data QC; population structure and cryptic relatedness; GWAS and PheWAS analysis; and data access and dissemination. All supplementary figures are inlined in the appropriate sections.

Supplementary Note 1

Discussion of noteworthy findings from 15 previously well-studied benchmark diseases.

Supplementary Note 2

A list of all FinnGen working-group members and their affiliations.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Kurki, M.I., Karjalainen, J., Palta, P. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023). https://doi.org/10.1038/s41586-022-05473-8

- Received: 10 January 2022
- Accepted: 21 October 2022
- Published: 18 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05473-8

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

Mono- and biallelic variant effects on disease at biobank scale

- H. O. Heyne
- J. Karjalainen
- M. J. Daly

Nature Article Open Access 18 Jan 2023

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05473-8

- Article
- Open Access
- Published: 18 January 2023

Mono- and biallelic variant effects on disease at biobank scale

- H. O. Heyne ,
- J. Karjalainen,
- K. J. Karczewski,
- S. M. Lemmelä,
- W. Zhou,
- FinnGen,
- A. S. Havulinna,
- M. Kurki,
- H. L. Rehm,
- A. Palotie &
- …
- M. J. Daly

Show authors

Nature **volume 613**, pages 519–525 (2023)

- 5746 Accesses
- 172 Altmetric
- Metrics details

Abstract

Identifying causal factors for Mendelian and common diseases is an ongoing challenge in medical genetics[1]. Population bottleneck events, such as those that occurred in the history of the Finnish population, enrich some homozygous variants to higher frequencies, which facilitates the identification of variants that cause diseases with recessive inheritance[2,3]. Here we examine the homozygous and heterozygous effects of 44,370 coding variants on 2,444 disease phenotypes using data from the nationwide electronic health records of 176,899 Finnish individuals. We find associations for homozygous genotypes across a broad spectrum of phenotypes, including known associations with retinal dystrophy and novel associations with adult-onset cataract and female infertility. Of the recessive disease associations that we identify, 13 out of 20 would have been missed by the additive model that is typically used in genome-wide association studies. We use these results to find many known Mendelian variants whose inheritance cannot be adequately described by a conventional definition of dominant or recessive. In particular, we find variants that are known to cause diseases with recessive inheritance with significant heterozygous phenotypic effects. Similarly, we find presumed benign variants with disease effects. Our results show how biobanks, particularly in founder populations, can broaden our understanding of complex dosage effects of Mendelian variants on disease.

Main

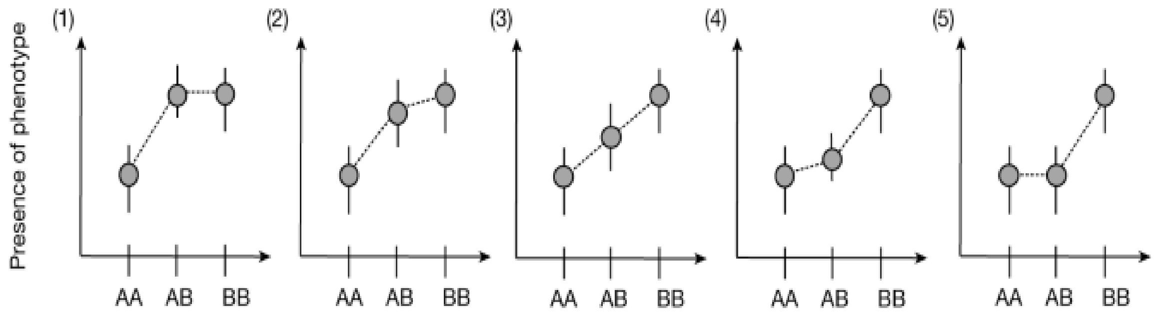
Rare genetic variants with large effects on disease can have direct implications for the development of potential treatments[1]; however, studying their effects comprehensively requires large, broadly phenotyped samples[4]. Identifying variants that influence disease risk only in the homozygous state (recessive inheritance) is particularly challenging, as the square of variant frequencies means that the homozygous state is often exceedingly rare. By contrast, in populations that have encountered a recent reduction in population size, certain founder diseases with recessive inheritance are present at higher frequencies. The Finnish population has experienced such bottleneck events and has been, historically, relatively isolated from other European populations[5]. As a result, the Finnish population is characterized by higher rates of DNA stretches with a common origin[6,7] carrying

particular sets of genetic variants. This leads to higher rates of homozygosity, and increases the chance occurrence of pathogenic variants in a homozygous state that lead to diseases with recessive inheritance. In consequence, there is an enrichment of 36 specific Mendelian genetic diseases such as congenital nephrotic syndrome, Finnish type (CNF)[8] in certain areas of Finland today that show mostly recessive inheritance. These 36 (at present) 'founder diseases' are referred to as the 'Finnish disease heritage'[2,9]. Analogous founder diseases in other populations include Tay-Sachs disease and Gaucher disease in Ashkenazi Jewish individuals[10]; Hermansky–Pudlak syndrome in Puerto Rican individuals[11]; and autosomal recessive spastic ataxia of Charlevoix–Saguenay and Leigh syndrome, French Canadian type in French Canadian individuals[12,13]. Populations that have undergone recent bottlenecks are also characterized by an excess of mildly deleterious variants, which are derived from rare variants that stochastically increased in frequency after a bottleneck event[4]. This has been previously shown in Finland through an excess of potentially deleterious probable loss-of-function (pLoF) variants at lower to intermediate frequencies (around 0.5%–5%)[3,14]. The higher allele frequencies of deleterious founder variants increases the statistical power for detecting disease associations. Isolated populations have thus been successfully used to map disease genes for decades[5,15,16].

Genetic variants can have different effects on disease when in a monoallelic state (only one allele carries the variant; heterozygous) or a biallelic state (both alleles carry the variant; homozygous) (see Fig. 1). For common genetic variants, early genome-wide association studies (GWASs) found that additive models captured most genotype–phenotype associations, including those with non-additive (also called dominance) effects[17]. The vast majority of GWASs were therefore conducted with additive models only. By contrast, Mendelian disease variants are rarely described as additive (or equivalently, as semidominantly inherited[18,19]). Even for well-known examples of semidominant inheritance, such as the *LDLR* gene, which is associated with familial hypercholesteremia, diseases caused by monoallelic and biallelic variants are listed as separate recessive and dominant conditions in standard databases[20] (for example, ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) and the Online Mendelian Inheritance in Man (OMIM) database (https://www.omim.org/)). Although this nomenclature is limiting, for many

Mendelian disease variants, monoallelic and biallelic phenotypes do have distinct features. For example, pathogenic variants in *ATM* cause cancer predisposition in a monoallelic and in a biallelic state, but ataxia telangiectasia only in a biallelic state. Genetic variants can also have distinct molecular effects that are differently inherited. In *SERPINA1*, for example, the same variant causes a dominantly inherited liver phenotype with a gain-of-function mechanism but a recessively inherited lung phenotype with a loss-of-function mechanism[21].

Fig. 1: Schema of different effect sizes of monoallelic (heterozygous) versus biallelic variant states.



Inheritance (rare disease)	Dominant	Dominant, with reduced penetrance and/or variable expressivity in heterozygotes	Semidominant	Recessive, with rare expressing heterozygotes	Recessive
Genotype dosage effect	Biallelic effect equals heterozygous effect (nonlinear, threshold)	Biallelic effect marginally larger than heterozygous effect (nonlinear)	Biallelic effect double of heterozygous effect (linear, additive)	Biallelic effect exceeding heterozygous effect (nonlinear)	Biallelic effect, no heterozygous effect (nonlinear, threshold)
Remark			Standard GWAS model		

A is the wild-type and B is the mutant allele. We distinguish five main scenarios that are associated with different modes of inheritance used in rare-disease genetics (first row of the table at the bottom). In rare-disease genetics, the phenotypes associated with the mono- and the biallelic state in scenarios 2, 3 and 4 are usually viewed as distinct disease entities, with the monoallelic phenotype regarded as dominantly inherited and the biallelic phenotype, which is usually more severe, regarded as recessively inherited. In the schema, we focus on autosomal inheritance and do not show overdominant or underdominant inheritance (rare outside the HLA region).

A perfectly linear additive genetic architecture (scenario 3) is also described, in which no dominance effect contributes to phenotypic variation.

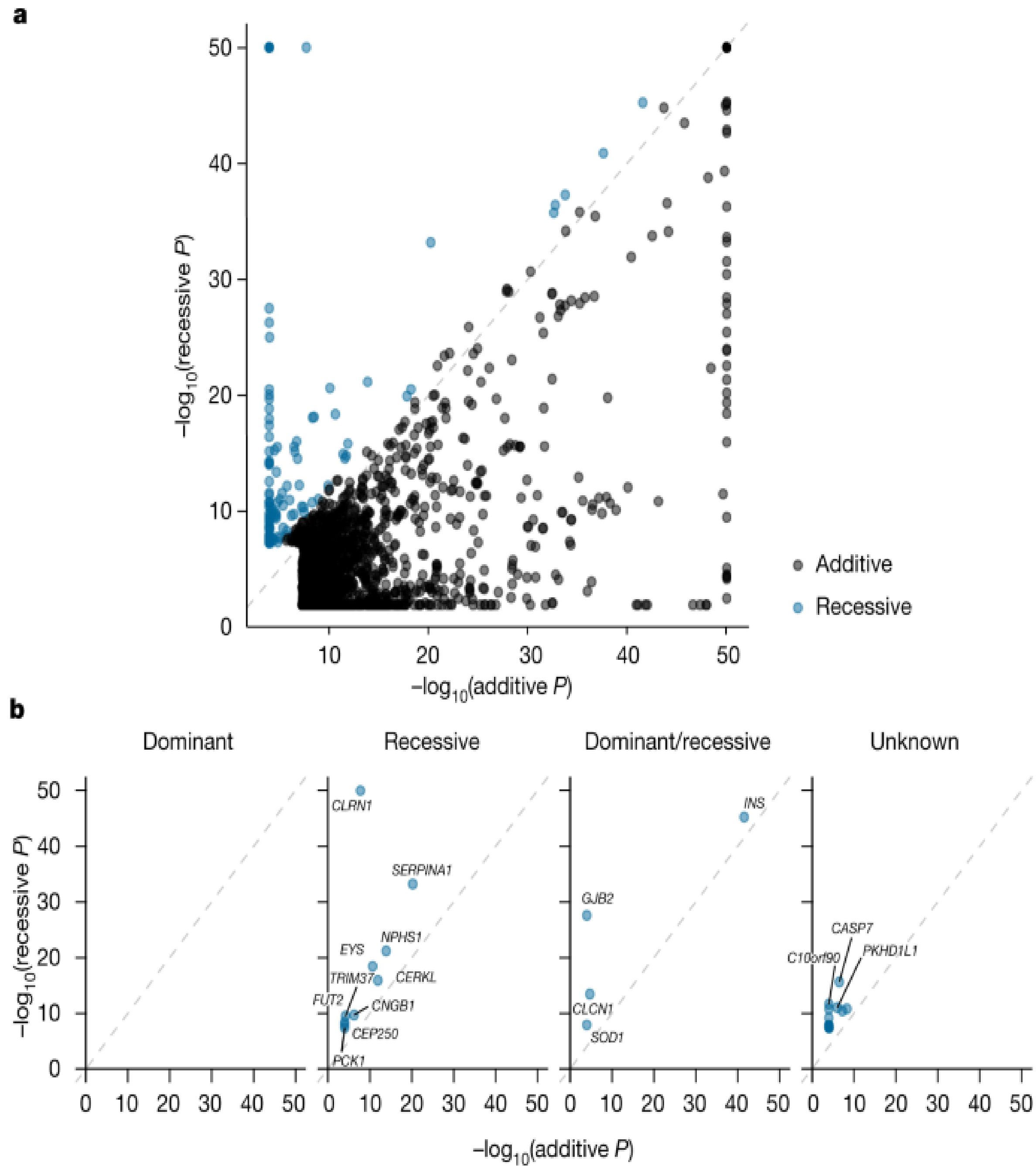
In this study, we analyse the effects of coding variants on 2,444 disease phenotypes using data from nationwide electronic health records of 176,899 Finnish individuals from the FinnGen project[22]. Participants are largely recruited through hospital biobanks and the data are thus enriched for individuals with diseases across the clinical spectrum. The phenotypes are derived from national healthcare registries collected over more than 50 years. In addition to the standard additive GWAS model[23], we systematically search for recessive associations. This enables us to examine two related questions of interest to both Mendelian and quantitative genetics communities. First, we investigate the potential benefit of searching for recessive associations in GWASs. Second, we consider the broad phenotypic consequences of Mendelian disease variants in the heterozygous as well as the homozygous state, and highlight how the current nomenclature could be improved to more precisely describe the complex inheritance of Mendelian variants.

Recessive disease associations

In light of the global enrichment of deleterious variants in Finland[3,14] and the well-described Finnish disease heritage[2,9], we analysed data from the large population cohort gnomAD[24] and found that a set of variants that are known to cause disease with recessive inheritance are enriched in the Finnish population (Extended Data Fig. 1 and Supplementary Note 1). We thus saw the opportunity to identify novel recessive associations in our Finnish dataset. We performed a phenome-wide association study (pheWAS) to search for the effects of coding variants on 2,444 disease phenotypes in 176,899 Finnish individuals. When using the conventional GWAS model assuming additive genotype effects (see also scenario 3 in Fig. 1) in 82,647 variants, we found 1,788 significant associations ($P < 5 \times 10^{-8}$) for 445 coding variants in 305 genes[22]. For 44,370 coding variants with 5 or more homozygous individuals, we also investigated whether any of them had disease effects in the homozygous state (recessive GWAS model; see also scenario 5 in Fig. 1 and Supplementary Tables 1 and 2). We identified 124

associations—involving 39 unique variants—where the recessive model fitted substantially better than the additive model (recessive *P* value two orders of magnitude smaller than the additive GWAS *P* value; Fig. 2), corresponding to 31 unique loci (Fig. 2b, Supplementary Table 3 and Supplementary Fig. 1). Our simulations (Supplementary Note 2, Extended Data Fig. 2 and Supplementary Table 4) supported this as a suitable way to identify recessive associations. The simulations also showed how a recessive model can uncover associations that are missed by a conventional GWAS model—particularly for rare variants (minor allele frequency (MAF) < 0.05). This is consistent with more than half of our recessive associations not being genome-wide significant in the additive GWAS model (Table 1). Another 93 associations (of 73 unique variants) were identified in which the recessive *P* value was smaller than the additive *P* value (Supplementary Table 1). Those likely included non-additive effects as a simulated additive effect only showed a smaller recessive *P* value than additive *P* value in around 2 in 100,000 simulations.

Fig. 2: *P* values of the additive versus the recessive GWAS model of all genome-wide-significant variant–disease associations.

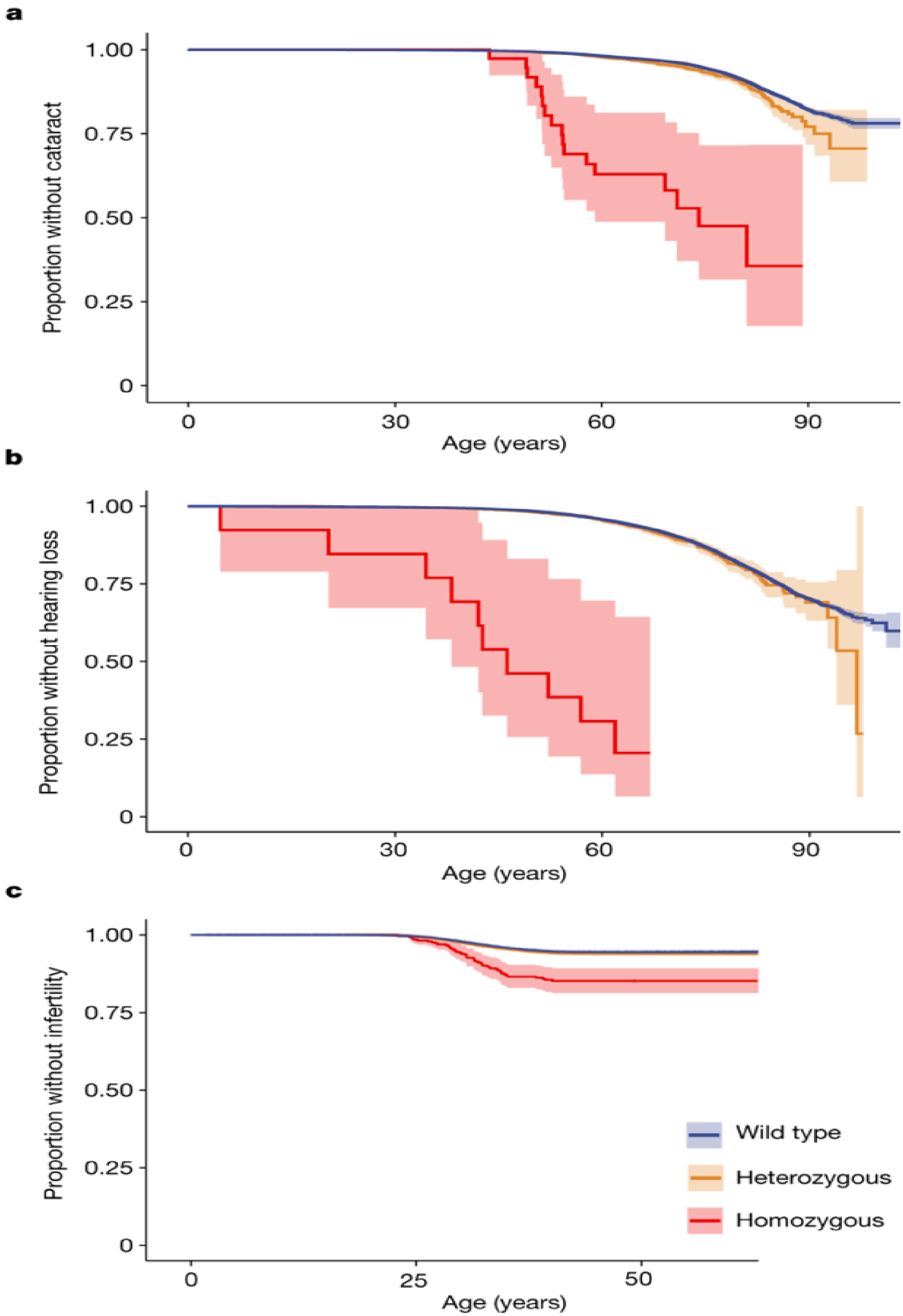


Associations in which the P value of the recessive model is two orders of magnitude lower than that of the additive model are shown in blue (category: 'recessive'); all other associations are shown in black. **a**, All associations. **b**, Recessive associations, broken down by known inheritance modes of the respective disease gene (source: OMIM). Independent loci (considering adjacent variants with $r^2 > 0.25$ associated with the same (parent) trait as one locus). For clarity, $-\log_{10} P$ values are capped at 50.

Table 1 Recessive associations

Of the 31 unique loci, 13 were in known OMIM genes (Fig. 2), all of which were previously described with recessive disease inheritance. Of these, 9 of the 13 variants were known disease-causing variants[20] (6 pathogenic or likely to be pathogenic ('likely pathogenic'); 3 likely pathogenic by at least one submitter) and had large effects on disease, as expected. Among the remaining 18 variants in non-OMIM genes, we highlight *CASP7* (adult-onset cataract) and *EBAG9* (female infertility) (Fig. 3 and Supplementary Note 3). We investigated the effect of the *EBAG9* variant on fertility in 106,732 women born between 1925 and 1975 (thus with an approximately complete reproductive period) in our FinnGen data. We found that individuals homozygous for the *EBAG9* variant had fewer children (1.7 ± 1.3 (mean ± s.d.)) than did wild-type individuals (2.0 ± 1.3, Wilcoxon rank test $P = 0.0002$), and had their first child at a later age (Extended Data Fig. 3), supporting the association of this variant with infertility. The six common (allele frequency greater than 10%) variants with recessive associations all overlapped known associations in the GWAS catalogue; however, for four of the six, a non-additive model of association had not been previously documented (Supplementary Table 5; variants in linkage disequilibrium with $r^2 > 0.1$; ref. [25]). Among the less-common variants, genes *GJB2* and *EYS* each contained two significantly associated likely pathogenic variants. For those and other genes, we thus identified compound heterozygous effects; that is, biallelic effects of two different pathogenic variants on separate alleles (see Extended Data Fig. 4 and Supplementary Note 4).

Fig. 3: Age at first diagnosis of variants with recessive disease associations.


a
Proportion without cataract
1.00
0.75
0.50
0.25
0
0
30
60
90
Age (years)
b
Proportion without hearing loss
1.00
0.75
0.50
0.25
0
0
30
60
90
Age (years)
c
Proportion without infertility
1.00
0.75
0.50
0.25
0
0
25
50
Age (years)
Wild type
Heterozygous
Homozygous

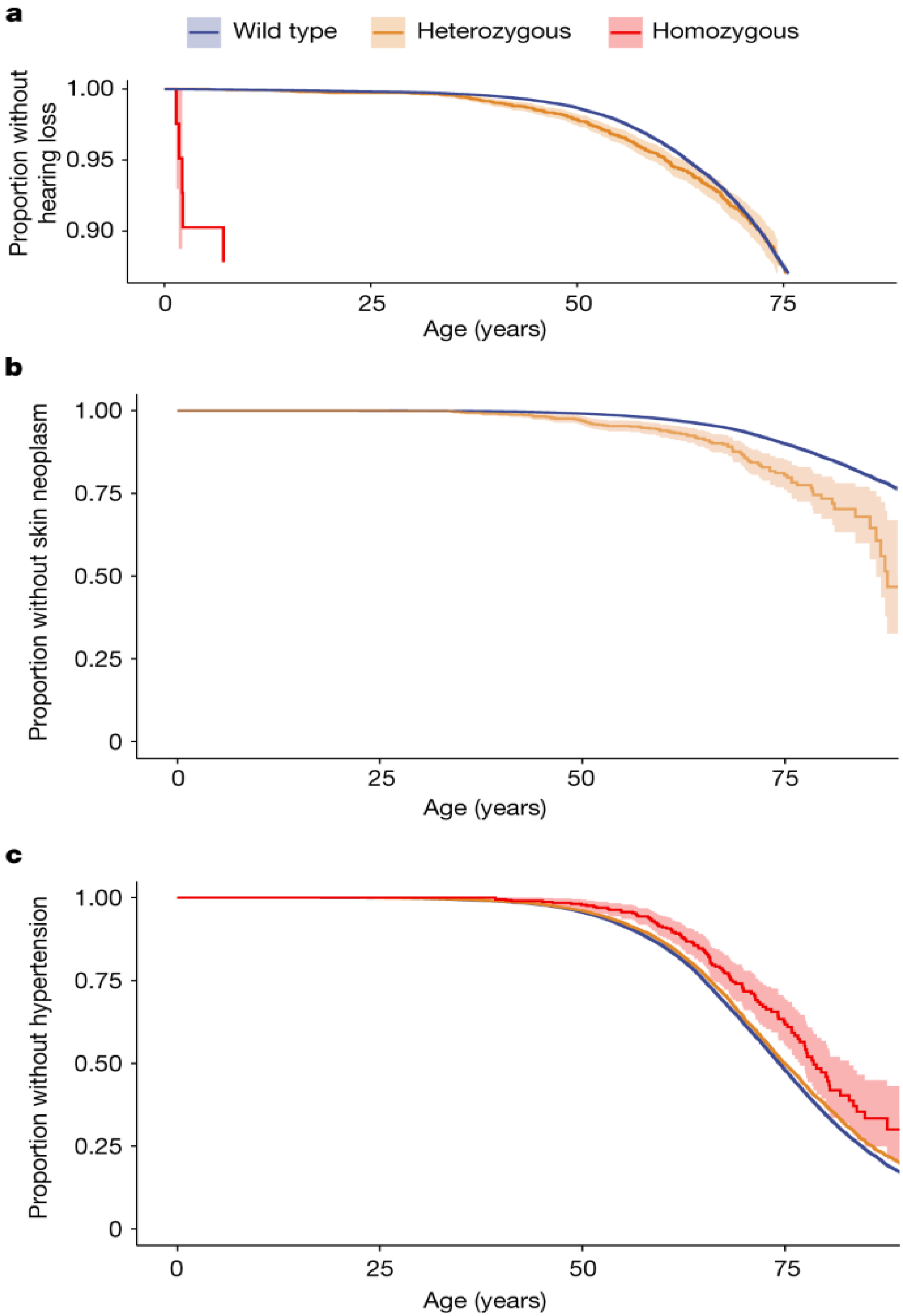
Data are shown as survival plots. **a**, Missense variant in *CASP7* associated with cataract; not previously described ($P = 2.5 \times 10^{-16}$); $n = 176{,}899$. **b**, Missense variant in *C10orf90* associated with hearing loss ($P = 2.2 \times 10^{-12}$); only recently described[53]; $n = 176{,}899$. **c**, Intronic variant in *EBAG* associated with female infertility ($P = 1.6 \times 10^{-11}$); not previously described; $n = 110{,}361$ female individuals. Survival curves of wild-type individuals are coloured in blue, heterozygotes in yellow and homozygotes in red. The 95% confidence intervals of the point estimates are shaded in light blue, light yellow or light red.

We next sought to validate any recessive associations in the UK Biobank (UKBB)[26] and in the FinnGen data freeze R6, an updated larger version of our original dataset. Owing to the isolation of the Finnish population[2], most of our variants with outsized homozygous effects were Finnish-enriched (allele frequency more than two times higher than that in Europeans who are not Finnish, Swedish or Estonian). Thus, in total, only 13 out of 31 variants were suitable for validation in the UKBB, 8 of which had nominally significant recessive associations to related phenotypes in the UKBB (Extended Data Fig. 5 and Supplementary Tables 6 and 7). We excluded seven variants with lower homozygote case counts in R6 compared to R4 as those may be due to technical differences between data freezes. Of the remaining 24 variants, 18 remained genome-wide significant in FinnGen R6. In total, we validated 20 recessive associations in FinnGen R6 and/or the UKBB (Table 1), including 4 novel associations. Recessive associations that could not be validated are listed in Supplementary Tables 1 and 3. Most homozygous individuals also had a substantially earlier onset of disease than did wild-type individuals affected with the same diseases, providing additional evidence for the effect of these variants on disease (Extended Data Fig. 6).

We next investigated whether any of the variants with outsized homozygous effects also had subtle heterozygous effects on the same diseases, by excluding homozygous and compound heterozygous individuals from the analysis. We found variants with nominally significant heterozygous effects in *SERPINA1*, *NPHS1* and *CASP7* (see Table 1). In addition, we confirmed a previously hypothesized[27] heterozygous effect of a pLoF variant in *GJB2* on hearing loss in the larger R6 replication dataset (Fig. 4b; $P = 0.02$, $\beta = 0.11$).

We thus classified the inheritance of these variants as recessive, with rare expressing heterozygotes (Fig. 1).

Fig. 4: Age at first diagnosis of known disease-associated variants.


a
Wild type
Heterozygous
Homozygous
Proportion without hearing loss
1.00
0.95
0.90
0
25
50
75
Age (years)
b
Proportion without skin neoplasm
1.00
0.75
0.50
0.25
0
0
25
50
75
Age (years)
c
Proportion without hypertension
1.00
0.75
0.50
0.25
0
0
25
50
75
Age (years)

Data are shown as survival plots. **a**, Known likely pathogenic variant (known recessive inheritance) in *GJB2* associated with hearing loss also in a heterozygous state ($P = 0.02$). The *y* axis is cut at 0.9 for clarity. **b**, Known likely pathogenic variant in *XPA* associated with skin cancer ($P = 8 \times 10^{-11}$). In a homozygous state, this variant causes xeroderma pigmentosum with childhood-onset skin cancer[40]. **c**, Likely benign missense variant in *DBH* protects from hypertension ($P = 5.2 \times 10^{-13}$). (*DBH* is associated with the recessively inherited disease dopamine beta-hydroxylase deficiency, which is characterized by severe hypotension[31]). **a** and **b** show R4 data ($n = 176,899$); **c** shows R6 data ($n = 234,553$). Survival curves of wild-type individuals are coloured in blue, heterozygous individuals in yellow and homozygous individuals in red. The 95% confidence intervals of the point estimates are shaded in light blue, light yellow or light red.

Effects of known disease variants

Public databases of variants such as ClinVar[20], the largest, are central for routine clinical genetics but—as with many research community efforts—can include errors. We investigated whether pheWASs of 2,444 disease phenotypes could potentially enhance the interpretation of some ClinVar variants. Specifically, such efforts could support ClinVar submissions that are seen in individual patients with statistically robust associations. We first cross-validated our own results, and found the disease associations for the most frequent likely pathogenic variants across a wide range of phenotypes as expected (see Supplementary Note 5, Supplementary Fig. 1 and Supplementary Table 8). In addition, we provide examples in which our data can help to verify or falsify previously described disease associations or identify novel associations (Fig. 4b). We then examined the collective association of several groups of ClinVar variants, as for many rare variants we were only powered to find their effects on disease with moderate significance (Methods and Extended Data Fig. 7). As anticipated, we found global disease associations for likely pathogenic variants in genes that are described to cause disease with dominant inheritance (classification: OMIM). However, we also found global disease associations for variants that are listed as benign or likely to be benign ('likely benign'). Likely benign variants are defined as 'not implicated in monogenic disease'[28], and

are often considered to be neutral[29]. However, 16 likely benign variants were likely causally associated with disease in our data (using statistical fine-mapping[30]; see Supplementary Table 9 and Supplementary Note 6). None of them had Mendelian effect sizes and would thus not cause monogenic disease. Rather, they moderately increased disease risk or protected from phenotypes that are mostly similar to the Mendelian phenotypes that are associated with the same gene. Of these benign variants, we highlight a variant in the gene *DBH*. *DBH* is a gene associated with dopamine beta-hydroxylase deficiency (inheritance: recessive), which is characterized by severe hypotension[31]. We observed a likely benign missense variant in *DBH* that conveyed protection from hypertension (see Fig. 4c), a plausible finding given the association of *DBH* with hypotension. The variant is also an example for a Finnish-enriched variant (allele frequency of 0.05, which is 22 times higher than that in Europeans who are not Finnish, Swedish or Estonian).

Variants with monoallelic and biallelic effects

Because we observed unexpected disease effects from heterozygous variants in genes with reported recessive inheritance, we sought to further contrast the effects of variants with their previously described modes of inheritance. We found multiple coding variants with effects that did not match their annotated inheritance. These include known variants in *CHEK2*, *JAK2*, *TYR*, *OCA2* and *MC1R* (previously described as dominant inheritance) that have additive effects on cancer and cancer-related phenotypes (Supplementary Fig. 2). Similarly, we highlight one variant in *SCN5A* that was previously associated with severe cardiac-arrhythmia-like sick sinus syndrome[32] in a biallelic state, which we confirm in our data (Fisher's exact test, $P = 9 \times 10^{-4}$; odds ratio (OR) = 48 (95% confidence interval: 6–319)). In a heterozygous state, however, that same *SCN5A* variant protects from cardiac arrhythmia in FinnGen ($\beta = -0.48$, $P = 2 \times 10^{-8}$, posterior inclusion probability = 0.996 (ref. [30]), indicating probable causality), including atrial fibrillation ($\beta = -0.62$, $P = 7 \times 10^{-7}$). We could replicate this association in the UKBB[26] ($\beta = -0.39$, $P = 0.04$). In line with the subtle heterozygous effects of individual variants, we found global disease effects of 203 likely pathogenic variants in disease genes with recessive inheritance (see

Extended Data Fig. 7 and Methods). Both simulations and reanalysis after excluding variants with homozygotes from the global analysis indicated that modest heterozygous effects are likely to contribute to this signal. In disease genes with known recessive and/or dominant inheritance, we found 79 additional coding variants (not likely pathogenic) with effects on disease that are likely to be additive (additive $P <$ recessive P) and causal[30]. Of these 79 variants, 11 had Mendelian effect sizes (OR > 3). In summary, we identify several variants that have disease effects in a heterozygous and a homozygous state in known Mendelian disease genes. The modes of inheritance of these variants cannot be described with the usual labels of recessive and dominant. Our data thus indicate the need for a nomenclature that integrates Mendelian and complex genetic effects on diseases. We outline a suggestion in Fig. 1.

Discussion

A subset of disease-causing variants is enriched to unusually high frequencies in populations with a history of recent bottlenecks, such as the Finnish population. This applies particularly to homozygous variants that cause disease with recessive inheritance. The FinnGen cohort, given its size and its inclusion of broad medical phenotypic data over the lifespan of an individual, is well-powered for discovering novel alleles with large effects in homozygous individuals and for studying the disease architecture of Mendelian variants.

We found multiple variants in known Mendelian disease genes[20] with large effects in homozygous individuals that have weaker but significant effects in heterozygous individuals for the same or closely related diseases, highlighting that describing their inheritance with only dominant and recessive labels does not adequately describe disease biology. Terms beyond recessive and dominant are, however, rarely used in clinical genetics. Although it has been estimated that semidominant inheritance is much more common than true dominant inheritance[33], phenotypes of homozygotes are frequently unknown because they are rare or too severe for an individual with the phenotype to survive to birth. The term semidominant inheritance is more frequently used in animal and plant genetics[34], in which mono- and biallelic effects can be more easily quantified and systematically studied.

Aggregating the evidence for pathogenicity across biallelic and monoallelic observations, rather than viewing these as separate disease entities, could benefit the interpretation of clinical variants, as well as providing a more accurate description of biology. Furthermore, we found that several variants that were previously described as likely to be benign are associated with disease. This is a noteworthy reminder that such likely benign variants are generally defined as not causing monogenic disease and should not necessarily be regarded as neutral[29].

Adding to an increased complexity of Mendelian inheritance, we detected modest heterozygous effects on disease in variants that are known to cause disease only in a homozygous state (recessive inheritance). This is in line with previous studies that found subtle heterozygous fitness effects of variants that are described as causing disease with recessive inheritance in mice[35], *Drosophila*[36] and humans[37]. We found heterozygous effects of such variants in the genes *NPHS1* (previously debated[38]), *SERPINA1* (known[21]) and *GJB2* (previously hypothesized[27]). As expected for recessive inheritance, the variants had an order-of-magnitude-larger effect size and, often, an earlier onset of disease in a homozygous than in a heterozygous state, thereby exceeding a linear additive model. We also found a heterozygous pLoF variant in *XPA* that increases susceptibility to adult-onset skin cancer. The effect was far larger than is usually found by GWASs and could thus provide valuable information on personal risk. Although a nominally significant association ($P = 0.01$) of a heterozygous pLoF variant in *XPA* with skin cancer was found in a previous study[39], we provide here the first—to our knowledge–definitive evidence. Homozygous pLoF variants in the gene *XPA* are known to cause a related phenotype, xeroderma pigmentosum, which is characterized by extreme vulnerability to UV radiation and childhood-onset skin cancer[40]. Long before these large-scale data became available, small heterozygous effects were found in variants that cause Mendelian disease with recessive inheritance[41], in some cases conferring an advantage against certain infectious diseases[42,43,44]. Similarly, we find that one variant in *SCN5A*, which was previously associated with severe cardiac arrhythmia such as sick sinus syndrome[32] in a biallelic state, protected from mild cardiac arrhythmia diseases in a heterozygous state in our data. Previous experimental data found a mild loss-of-function effect of this variant[32,45]. This is in line with a potential slowing of electrical

conduction in the heart in a few individuals who are heterozygous for the *SCN5A* variant[32], which could thus provide a protective mechanism against cardiac arrhythmia.

Of course, it is possible that this and other heterozygous effects we observe come from low-frequency variants in a compound heterozygous state that were not captured by our genotyping array. However, a different age of disease onset in heterozygous than in homozygous individuals suggests that that scenario is unlikely to explain most of the observed heterozygous effect. In addition, with population-specific exome sequencing and imputation we could account for the presence of additional pathogenic variants at frequencies greater than 0.2%. Additional limitations are the lack of more in-depth phenotypes, including symptoms that are not captured by International Classification of Diseases (ICD) codes, or serological or diagnostic tests missing subtle physiological differences between heterozygous and wild-type individuals. Furthermore, the Finnish population bottleneck leads to a lower number of rare variants, which limits new discoveries. We cannot thus investigate relatively common European Mendelian disease variants; for example, variants in *CFTR*, which cause cystic fibrosis[46].

We systematically investigated recessive associations of coding variants genome-wide with pheWAS at biobank scale. We could validate 20 loci (4 of them novel) that had large biallelic effects without, or with only nominally significant heterozygous effects. Novel associations included complex non-syndromic diseases with few or no previously described large-effect variants, such as adult-onset cataract (new disease gene: *CASP7*) and female infertility (new disease gene: *EBAG9*). Intuitively, these novel findings appeared in phenotypes (deafness, cataracts and infertility) for which Mendelian subtypes would not obviously be clinically distinguished from other common presentations, although—as in previous examples—an earlier age of onset in the cases of common diseases of ageing is seen. Biallelic associations of rare coding variants have been found in other population biobanks[47,48,49,50,51], but have not (to our knowledge) been investigated in a broad phenotype context in most studies, which lack the scale, the advantage of isolation and/or the high rates of homozygosity by descent. We suggest that searching for homozygous effects is most meaningful for coding and

structural variants as opposed to regulatory variants with only weak effects. This is consistent with the observations that dominance effects are very modest in GWASs of common variants in complex human traits[52], but that recessive inheritance in Mendelian disease is widespread.

In summary, our biobank-scale additive and recessive pheWAS of coding variants shows the benefit of including recessive scans in GWASs. We find known and novel biallelic associations across a broad spectrum of phenotypes such as retinal dystrophy, adult-onset cataract and female infertility that are missed by the standard additive GWAS model. As a related point, we find an underappreciated complexity of inheritance patterns of multiple Mendelian variants. Our study could thus provide a starting point for reconciling the variant-effect nomenclature of the conventionally separate but more-and-more overlapping fields of Mendelian and complex genetics.

Methods

Ethics and data access approvals

The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) approved the FinnGen study protocol HUS/990/2017. The FinnGen project is approved by the Finnish Institute for Health and Welfare (THL) (approval number THL/2031/6.02.00/2017; amendments THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018 and THL/283/6.02.00/2019), the Digital and Population Data Services Agency (VRK43431/2017-3 and VRK/6909/2018-3), the Social Insurance Institution (KELA) (KELA 58/522/2017, KELA 131/522/2018 and KELA 70/522/2019) and Statistics Finland TK-53-1041-17. The Biobank access decisions for FinnGen samples and data used in FinnGen data freeze 4 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71 and BB2019_7, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealis of Northern Finland_2017_1013, Biobank of Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland Biobank 1–2017 and Terveystalo Biobank STB 2018001.

Patients and control individuals in FinnGen provided informed consent for biobank research, according to the Finnish Biobank Act. Alternatively, separate research cohorts, collected before the Finnish Biobank Act came into effect (in September 2013) and before the start of FinnGen (August 2017) were collected on the basis of study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is HUS/990/2017. UK biobank data were accessed under protocol 31063.

Funding and partners

We acknowledge the participants and investigators of the FinnGen study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie, AstraZeneca UK, Biogen MA, Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech, Merck Sharp & Dohme, Pfizer, GlaxoSmithKline Intellectual Property Development, Sanofi US Services, Maze Therapeutics, Janssen Biotech, Novartis and Boehringer Ingelheim International. The following biobanks are acknowledged for delivering biobank samples to FinnGen: Arctic Biobank (https://www.oulu.fi/medicine/node/207208), Auria Biobank (https://www.auria.fi/biopankki) THL Biobank (https://thl.fi/en/web/thl-biobank), Helsinki Biobank (https://www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (https://ita-suomenbiopankki.fi/en/), Central Finland Biobank (https://www.sairaalanova.fi/en-US), Finnish Red Cross Blood Service Biobank (https://www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank (https://www.terveystalo.com/fi/yhtio/biopankki) and the Finnish Hematology Registry and Clinical Biobank (https://www.fhrb.fi). All Finnish Biobanks are members of the BBMRI infrastructure

(https://www.bbmri-eric.eu/national-nodes/finland/). Finnish Biobank Cooperative (FINBB; https://finbb.fi) is the coordinator of BBMRI-ERIC operations in Finland.

Coding variants in FinnGen, release 4

Greater haplotype sharing in the Finnish population facilitates the imputation of lower-frequency variants in array data with a population-specific reference panel down to frequencies below 0.0005 (ref. [54]). Genotypes were thus generated with arrays thereby enabling the large scale of the FinnGen research project. In summary, we investigated 82,647 coding variants (2,634 pLoF, 76,884 missense and 3,129 others) in the FinnGen project, release 4, 8/2019 in 176,899 Finnish individuals. We excluded the HLA region (chr. 6, 25 Mb–35 Mb). In 110,361 individuals, sex was imputed as female. We filtered to variants with INFO > 0.8 at a median allele frequency of 0.004 (minimum allele frequency 3×10^{-5}). Around half of the samples came from existing legacy collections, and the other half was from participants who were newly recruited to the FinnGen project. Samples were genotyped on custom microarrays and rare variants were imputed using a population-specific reference panel[54]. To calculate variant enrichment in Finnish individuals after a bottleneck event, we used as a general European reference point exomes from European samples in gnomAD 2.1.1, excluding those from Finland, Sweden and Estonia. Owing to large-scale migrations from Finland to Sweden in the 20th century, a substantial fraction of the genetic ancestry in Sweden is of recent Finnish origin, and the linguistically (and geographically) close population of Estonia is likely to share elements of the same ancestral founder effect. See ref. [22] or https://finngen.gitbook.io/documentation/ for a detailed description of data production and analysis.

GWAS searching for additive and recessive associations

We performed a GWAS on 2,444 disease end-points, investigating the effects of 82,647 coding variants with an additive and recessive model using the method SAIGE[23]. Covariates in FinnGen were age, sex, genotyping batch and the first 10 principal components of genotypes. To identify heterozygous effects, we performed a GWAS with an additive model after

excluding homozygous and compound heterozygous individuals, where possible (see Supplementary Note 4). In the recessive GWAS model we analysed the effects of homozygous alleles on disease phenotypes in comparison to wild-type and heterozygous alleles. The wild type is defined as the reference allele in FinnGen. The docker container finngen/saige:0.39.1.fg with all necessary software used to run SAIGE in additive or recessive mode can be found at the docker container library hub.docker.com. We replicated our genome-wide significant recessive associations in 234,553 individuals of FinnGen data freeze R6 and in 420,531 individuals with European ancestry in the UKBB using the same recessive model in SAIGE. GWAS covariates in UKBB were age, sex, age × sex, age^2, age^2 × sex and the first 10 principal components of genotypes. The genotype and phenotype files along with ancestry definitions, phenotype definitions and SAIGE null models were taken from the PAN UKBB project and are further described at https://pan.ukbb.broadinstitute.org. For additional information on and results from FinnGen data freeze R6 please search the indicated websites in the 'Data availability' section.

Annotating variant effects from ClinVar

We annotated variants from release 25 March 2020 of ClinVar[20]. For any variant included in the main tables, we rechecked current classifications in ClinVar and OMIM on 2 November 2021. We grouped variants into categories according to their 'ClinVar_ReviewStatus'. Our main categories were likely pathogenic (likely to be pathogenic or pathogenic; 311 variants), conflicting evidence (at least one submitter labelled a variant as likely pathogenic but at least one other submitter labelled it different from likely pathogenic; 298 variants) and likely benign (likely to be benign or benign; 10,948 variants). The ClinVar annotation labelling variants as likely benign was above average quality (1.6; 7 of the 16 likely benign top causal variants had a one-star and 9 of 16 a two-star review status in ClinVar) compared to an average 1.2 stars for all likely benign variants in ClinVar (range: zero to three stars). Other categories into which we grouped variants that we are not explicitly discussing in the manuscript were 'association' (26 variants), 'drug response' (59 variants), 'not provided' (141 variants), 'protective' (14 variants), 'risk factor' (74 variants) and 'VUS' (variant of unknown significance; 3,269 variants); see Supplementary Table 8.

Annotating inheritance mode from OMIM

We downloaded the OMIM catalogue of human genetic diseases (https://www.omim.org/) version 06/2019. From OMIM, we annotated genes implicated to cause disease with a recessive or dominant inheritance mode.

Global phenotype associations of ClinVar variant categories

We compared global disease phenotype associations of different ClinVar variant categories (likely benign, likely pathogenic or conflicting variants in genes with dominant or recessive inheritance) with phenotype associations of random intergenic variants. For a given variant category, we counted how many variants had at least one significant GWAS hit (2,444 phenotypes) below a given *P* value threshold. We then compared those to the number of top GWAS loci below the *P* value threshold of 1,000 random samples of intergenic variants. We calculated with empirical *P* values if any ClinVar variant categories had significantly more disease associations than random intergenic variants below respective *P* value thresholds. Allele frequency influences the power with which significant associations are identified. Therefore, we adjusted for allele frequency by sampling intergenic variants in 15 equal-sized bins that corresponded to the allele frequency of the variants under investigation. To account for linkage disequilibrium, we sampled intergenic variants from the same 3-Mb windows as variants in the respective gene set.

Age at first diagnosis

We compared age at first diagnosis of homozygous or heterozygous individuals compared to wild-type individuals, respectively, using Wilcoxon rank tests. For a few compound heterozygous variants (as indicated in the paper), we also performed survival analyses using age at first disease diagnosis as outcome using a Cox proportional hazard model with the same covariates that were also used in the GWAS (sex, age, genotyping batch and first 10 principal components).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All summary statistics described in this manuscript can be found in the Supplementary Information. All information for downloading the summary statistics of additive GWASs of FinnGen release 4 can be found at https://finngen.gitbook.io/documentation/v/r4/data-download. Information for downloading the FinnGen release 6 summary statistics can be found at https://finngen.gitbook.io/documentation/v/r6/data-download. You can learn more about accessing other FinnGen data here: https://www.finngen.fi/en/access_results. A full list of FinnGen endpoints for release 4 and release 6 is available at https://www.finngen.fi/en/researchers/clinical-endpoints. A table of coding variants' summary statistics (any additive association with P value $< 1 \times 10^{-4}$) has been included as a Supplementary Table 10. Individual-level genotypes and register data from FinnGen participants can be accessed by approved researchers through the Fingenious portal (https://site.fingenious.fi/en/) hosted by the Finnish Biobank Cooperative FinBB (https://finbb.fi/en/). Data release to FinBB is timed to the biannual public release of FinnGen summary results, which occurs 12 months after members of the FinnGen consortium can start working with the data.

Code availability

Please see https://finngen.gitbook.io/documentation/ for a detailed description of data production and analysis, including the code that was used to run analyses, and https://github.com/FINNGEN/ for further code repositories that were used to run analyses in FinnGen. R code to reproduce figures is available upon request.

References

1. Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

2. Peltonen, L., Jalanko, A. & Varilo, T. Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.* **8**, 1913–1923 (1999).

3. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).

4. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).

5. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* **1**, 182–190 (2000).

6. Kerminen, S. et al. Fine-scale genetic structure in Finland. *G3* **7**, 3459–3468 (2017).

7. Martin, A. R. et al. Haplotype sharing provides insights into fine-scale population history and disease in Finland. *Am. J. Hum. Genet.* **102**, 760–775 (2018).

8. Fuchshuber, A. et al. Presymptomatic diagnosis of familial steroid-resistant nephrotic syndrome. *Lancet* **347**, 1050–1051 (1996).

9. Polvi, A. et al. The Finnish disease heritage database (FinDis) update-a database for the genes mutated in the Finnish disease heritage brought to the next-generation sequencing era. *Hum. Mutat.* **34**, 1458–1466 (2013).

10. Ostrer, H. A genetic profile of contemporary Jewish populations. *Nat. Rev. Genet.* **2**, 891–898 (2001).

11. Wildenberg, S. C. et al. A gene causing Hermansky-Pudlak syndrome in a Puerto Rican population maps to chromosome 10q2. *Am. J .Hum. Genet.* **57**, 755–765 (1995).

12. Bouchard, J. P., Barbeau, A., Bouchard, R. & Bouchard, R. W. Autosomal recessive spastic ataxia of Charlevoix-Saguenay. *Can. J. Neurol. Sci.* **5**, 61–69 (1978).

13. Mootha, V. K. et al. Identification of a gene causing human cytochrome *c* oxidase deficiency by integrative genomics. *Proc. Natl Acad. Sci. USA* **100**, 605–610 (2003).

14. Locke, A. E. et al. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**, 323–328 (2019).

15. Macdonald, M. E. et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).

16. Moises, H. W. et al. An international two-stage genome-wide search for schizophrenia susceptibility genes. *Nat. Genet.* **11**, 321–324 (1995).

17. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

18. Byrnes, A. M. et al. Mutations in *GDF5* presenting as semidominant brachydactyly A1. *Hum. Mutat.* **31**, 1155–1162 (2010).

19. Wilkie, A. O. The molecular basis of genetic dominance. *J. Med. Genet.* **31**, 89–98 (1994).

20. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

21. Crystal, R. G. α1-Antitrypsin deficiency, emphysema, and liver disease. Genetic basis and strategies for therapy. *J. Clin. Invest.* **85**, 1343–1352 (1990).

22. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* https://doi.org/10.1038/s41586-022-05473-8 (2023).

23. Zhou, W. et al. Efficiently controlling for case–control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

24. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

25. Lin, S. H., Brown, D. W. & Machiela, M. J. LDtrait: an online tool for identifying published phenotype associations in linkage disequilibrium. *Cancer Res.* **80**, 3443–3446 (2020).

26. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

27. Chan, D. K. & Chang, K. W. GJB2-associated hearing loss: systematic review of worldwide prevalence, genotype, and auditory phenotype. *Laryngoscope* **124**, E34–E53 (2014).

28. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

29. Niroula, A. & Vihinen, M. How good are pathogenicity predictors in detecting benign variants. *PLoS Comput. Biol.* **15**, e1006481 (2019).

30. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. B* **82**, 1273–1300 (2020).

31. Robertson, D. et al. Isolated failure of autonomic noradrenergic neurotransmission.*N. Engl. J. Med.* **314**, 1494–1497 (1986).

32. Benson, D. W. et al. Congenital sick sinus syndrome caused by recessive mutations in the cardiac sodium channel gene (*SCN5A*). *J. Clin. Invest.* **112**, 1019–1028 (2003).

33. Wilkie, A. O. Dominance and recessivity. *eLS* https://doi.org/10.1038/npg.els.0005475 (2006).

34. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl Acad. Sci. USA* **109**, 14746–14753 (2012).

35. Amorim, C. E. G. et al. The population genetics of human disease: the case of recessive, lethal mutations. *PLoS Genet.* **13**, e1006915 (2017).

36. Burkhart, B. D., Montgomery, E., Langley, C. H. & Voelker, R. A. Characterization of allozyme null and low activity alleles from two natural populations of *Drosophila melanogaster*. *Genetics* **107**, 295–306 (1984).

37. Wright, C. F. et al. Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am. J. Hum. Genet.* **104**, 275–286 (2019).

38. Caridi, G. et al. Clinical features and long-term outcome of nephrotic syndrome associated with heterozygous *NPHS1* and *NPHS2* mutations. *Clin. J. Am. Soc. Nephrol.* **4**, 1065–1072 (2009).

39. Hirai, Y. et al. Increased risk of skin cancer in Japanese heterozygotes of xeroderma pigmentosum group A. *J. Hum. Genet.* **63**, 1181–1184 (2018).

40. Tanaka, K. et al. Analysis of a human DNA excision repair gene involved in group A xeroderma pigmentosum and containing a zinc-finger domain. *Nature* **348**, 73–76 (1990).

41. Vogel, F. Clinical consequences of heterozygosity for autosomal-recessive diseases. *Clin. Genet.* **25**, 381–415 (1984).

42. Gouagna, L. C. et al. Genetic variation in human *HBB* is associated with *Plasmodium falciparum* transmission. *Nat. Genet.* **42**, 328–331 (2010).

43. Pier, G. B. et al. *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature* **393**, 79–82 (1998).

44. Ayi, K., Turrini, F., Piga, A. & Arese, P. Enhanced phagocytosis of ring-parasitized mutant erythrocytes: a common mechanism that may explain protection against falciparum malaria in sickle trait and beta-thalassemia trait. *Blood* **104**, 3364–3371 (2004).

45. Butters, T. D. et al. Mechanistic links between Na+ channel (SCN5A) mutations and impaired cardiac pacemaking in sick sinus syndrome. *Circ. Res.* **107**, 126–137 (2010).

46. Kinnunen, S. et al. Spectrum of mutations in CFTR in Finland: 18 years follow-up study and identification of two novel mutations. *J. Cyst. Fibros.* **4**, 233–237 (2005).

47. Belbin, G. M. et al. Genetic identification of a common collagen disease in Puerto Ricans via identity-by-descent mapping in a health system. *eLife* **6**, e25060 (2017).

48. Gudbjartsson, D. F. et al. A frameshift deletion in the sarcomere gene *MYL4* causes early-onset familial atrial fibrillation. *Eur. Heart J.* **38**, 27–34 (2017).

49. Guindo-Martinez, M. et al. The impact of non-additive genetic associations on age-related complex diseases. *Nat. Commun.* **12**, 2436 (2021).

50. Belbin, G. M. et al. Toward a fine-scale population health monitoring system. *Cell* **184**, 2068–2083 (2021).

51. Barton, A. R., Hujoel, M. L. A., Mukamel, R. E., Sherman, M. A. & Loh, P.-R. A spectrum of recessiveness among Mendelian disease variants in UK Biobank. *Am. J. Hum. Genet.* **109**, 1298–1307 (2022).

52. Palmer, D. S. et al. Analysis of genetic dominance in the UK Biobank. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.15.456387 (2021).

53. Ivarsdottir, E. V. et al. The genetic architecture of age-related hearing impairment revealed by genome-wide association analysis. *Commun. Biol.* **4**, 706 (2021).

54. Kals, M. et al. Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses. Preprint at *bioRxiv* https://doi.org/10.1101/579201 (2019).

55. Grubb, R. Correlation between Lewis blood group and secretor character in man. *Nature* **162**, 933 (1948).

Acknowledgements

We thank all FinnGen participants for their contributions to the research. Patients and control individuals in FinnGen provided informed consent for biobank research, on the basis of the Finnish Biobank Act. Research cohorts collected before the Finnish Biobank Act were collected on the basis of study-specific consents and later transferred to the Finnish biobanks after approval by Fimea, the National Supervisory Authority for Welfare and Health. The FinnGen project, data release 4, was funded by Business Finland and 13 industry partners. Further information on funding and approvals can be found in 'Funding and partners' in the Methods and the full list of members of the FinnGen consortium is provided in the Supplementary Information. We thank L. Biesecker, A. L. George, J. Kosmicki and J. Krause for discussions and comments.

Funding

Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

Author information

Authors and Affiliations

1. Finnish Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland

   H. O. Heyne, J. Karjalainen, K. J. Karczewski, S. M. Lemmelä, W. Zhou, A. S. Havulinna, M. Kurki, A. Palotie & M. J. Daly

2. Digital Health Center, Hasso Plattner Institute for Digital Engineering, University of Potsdam, Potsdam, Germany

H. O. Heyne

3. Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA

   H. O. Heyne

4. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

   H. O. Heyne

5. Program for Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

   H. O. Heyne, J. Karjalainen, K. J. Karczewski, W. Zhou, M. Kurki, H. L. Rehm, A. Palotie & M. J. Daly

6. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

   J. Karjalainen, K. J. Karczewski, W. Zhou, M. Kurki, A. Palotie & M. J. Daly

7. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

   J. Karjalainen, K. J. Karczewski, W. Zhou, M. Kurki, H. L. Rehm, A. Palotie & M. J. Daly

8. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

   J. Karjalainen, K. J. Karczewski, W. Zhou, M. Kurki, H. L. Rehm, A. Palotie & M. J. Daly

9. Finnish Institute for Health and Welfare, Helsinki, Finland

   S. M. Lemmelä & A. S. Havulinna

10. Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA

A. Palotie

Consortia

FinnGen

- H. O. Heyne
- , J. Karjalainen
- , S. M. Lemmelä
- , W. Zhou
- , A. S. Havulinna
- , M. Kurki
- , A. Palotie
- & M. J. Daly

Contributions

H.O.H. wrote the manuscript and generated the figures. H.O.H. and J.K. performed analyses. K.J.K., S.M.L., A.S.H., W.Z., M.K. and H.L.R. contributed data and research tools. A.P. and M.J.D. supervised the study. H.O.H. and M.J.D. conceived the study. All authors listed under FinnGen contributed to the generation of the primary data of the FinnGen data release 4. All authors reviewed the manuscript.

Corresponding authors

Correspondence to H. O. Heyne or M. J. Daly.

Ethics declarations

Competing interests

A.P. is a member of the Pfizer Genetics Scientific Advisory Panel. M.J.D. is a founder of Maze Therapeutics. The remaining authors declare no competing interests.

Peer review

Peer review information

Nature thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Variants known to cause disease with recessive inheritance are found at a higher MAF in Finnish Europeans than in other Europeans.

We show MAF of 2,419 unique likely pathogenic variants (source: ClinVar[20]) in 10,824 individuals from FIN and 10,824 from NFE populations (source: gnomAD[24]). **a**, All 2,419 variants (violin plot). **b**, All 2,419 variants (histogram). In **c** we subset to 133 variants in Finnish disease heritage genes[2,9]. Violin plots are scaled to have the same area. Box plots within violins show the 1st, 2nd and 3rd quartiles of the MAF distribution; whiskers maximally extend to 1.5 interquartile range.

Extended Data Fig. 2 Simulations of recessive and additive effects at different MAFs.

Here, we generated genotype counts of wild types, heterozygotes and mutant homozygotes in 200,000 individuals of a variant with an allele frequency of

0.01 following Hardy Weinberg Equilibrium (R library: HardyWeinberg) and random controls and cases of a disease with a prevalence of 0.05 (see also Supplementary Note 2). In **a** we simulate a recessive association. Here, we set the probability of homozygotes to develop the disease to 5x compared to wild type and the heterozygous effect to 1 (= no effect). In this histogram we show on the x-axis, the log10 p-value of the recessive model - the log10 p-value of the additive model (method: logistic regression). In **b** we simulate an additive association. Here, we set the homozygous effect to 1.5x and the heterozygous effect to 2.25. In this histogram we show on the x-axis, the log10 p-value of the additive model - the log10 p-value of the recessive model (method: logistic regression).

Extended Data Fig. 3 *EBAG* is associated with female infertility.

a, Comparing number of offspring in 147,061 women who are wild type (n = 131,141), heterozygous (n = 15,455) or homozygous (n = 465) for the *EBAG9* variant. Among 7,980 women with children and diagnosed with infertility 71 *EBAG9* homozygotes (**b**) had fewer children and (**c**) had their first child significantly later (see Supplementary Note 5).

Extended Data Fig. 4 Age at first disease diagnosis of variant carriers in *GJB2* (survival plot).

Wt, wild type; het, heterozygous; hom, homozygous; comp-het, compound heterozygous; GT, genotype. Genotypes of a known pathogenic missense and pLoF variant in *GJB2* associated with hearing loss. Comp-het carry both the pathogenic missense and pLoF variant on different alleles.

Extended Data Fig. 5 Replications with UKBB data.

We replicated the 31 recessive associations in FinnGen in the UKBB[26] with a recessive model in SAiGE[23]. Of the 31 variants, 13 had ≥ 5 homozygotes in the UKBB of which 8 had a significant recessive p-value in UKBB for the same/similar phenotype. Here, we show recessive betas in FinnGen versus UKBB (betas set to 5, when beta > 5). Plot size corresponds to MAF in gnomAD non-Finnish Europeans and plot colour red means the recessive p-value in UKBB is < 0.05.

Extended Data Fig. 6 Age at disease onset.

This forest plot shows the median age at first diagnosis for each variant with recessive associations in our FinnGen data. P-values indicate differences in disease onset between respective homozygous or heterozygous compared to wild-type carriers (Wilcoxon rank tests). Bars represent the first and third quartile of age at first diagnosis. Only variants with more than 5 affected homozygotes are shown. The y-axis lists gene-disease associations. Homozygotes had significantly earlier (or later for a known homozygous protective variant in *FUT2*[55]) disease onset for 7/31 variants than wild types (p-value < 0.0016 with Bonferroni correction for 31 tests, 14/31 tests with nominal p-value < 0.05, Wilcoxon rank test.).

Extended Data Fig. 7 Global disease associations of variant categories.

Variants previously described as disease-causing (ClinVar likely pathogenic or conflicting variants[20]) but also ClinVar likely benign variants are globally associated with disease phenotypes compared with randomly sampled intergenic variants matched to the same minor allele frequency in 15 bins. Disease-causing variants in OMIM genes that were described with only dominant inheritance, as well as genes with only recessive inheritance were globally disease-associated.

Supplementary information

Supplementary Information

This file contains Supplementary Notes, Supplementary Tables 5–7 and 9, Supplementary References and a list of FINNGEN members.

Reporting Summary

Supplementary Figure 1

Longitudinal survival curves showing disease onset of homozygous, heterozygous and wildtypes of variants with recessive associations in FinnGen.

Supplementary Figure 2

Longitudinal survival curves showing disease onset of homozygous, heterozygous and wildtypes of known likely pathogenic or conflicting disease variants.

Peer Review File

Supplementary Table 1

Any coding variant association in the additive or recessive model with $P < 10^{-7}$ (all available disease end-points).

Supplementary Table 2

A more detailed version of Table 1. Recessive associations. Only the most significant variant is shown per locus. hom, homozygotes in FinnGen; R6, FinnGen data release 6 (n = 234,553); UKBB, UK biobank (n = 420,531).

Supplementary Table 3

Summary statistics of coding variants with any recessive association with $P < 10^{-7}$ (all available disease end-points, associations with $P < 0.01$).

Supplementary Table 4

Simulation results of genotype and phenotype data to assess the output of our recessive and additive model in the presence of a strictly recessive or additive association.

Supplementary Table 8

Most significant disease associations of all 15,140 coding variants annotated in ClinVar.

Supplementary Table 10

Summary statistics of all coding variants in FinnGen release 4 (any additive association with $P < 1 \times 10^{-4}$).

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Heyne, H.O., Karjalainen, J., Karczewski, K.J. *et al.* Mono- and biallelic variant effects on disease at biobank scale. *Nature* **613**, 519–525 (2023). https://doi.org/10.1038/s41586-022-05420-7

- Received: 28 April 2021
- Accepted: 06 October 2022

- Published: 18 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05420-7

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

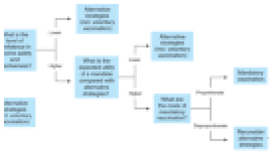
Provided by the Springer Nature SharedIt content-sharing initiative

FinnGen provides genetic insights from a well-phenotyped isolated population

- Mitja I. Kurki
- Juha Karjalainen
- Aarno Palotie

Nature Article Open Access 18 Jan 2023

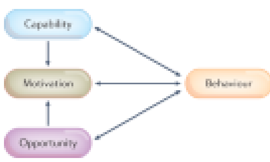
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05420-7



Balancing incentives and disincentives for vaccination in a pandemic

06 September 2021

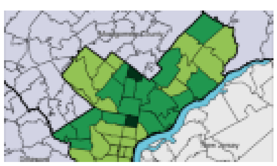
Julian Savulescu, Jonathan Pugh & Dominic Wilkinson



Encouraging vaccine uptake: lessons from behavioural science

27 July 2022

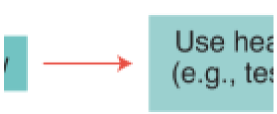
Susan Michie



A citywide experiment testing the impact of geographically targeted, high-pay-off vaccine lotteries

01 September 2022

Katherine L. Milkman, Linnea Gandhi, … Kevin G. Volpp



How kindness can be contagious in healthcare

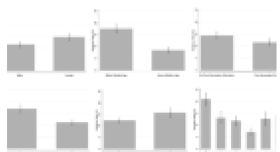
14 June 2021

Weiming Tang, Dan Wu, … Joseph D. Tucker

Rather than inducing psychological reactance, requiring vaccination strengthens intentions to vaccinate in US populations

21 October 2021

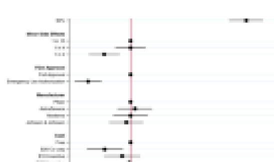
Dolores Albarracin, Haesung Jung, … Jessica Fishman



Characteristics associated with COVID-19 vaccine hesitancy

20 July 2022

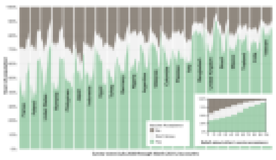
Liyousew G. Borga, Andrew E. Clark, … Anthony Lepinteur



Public attitudes toward COVID-19 vaccination: The role of vaccine attributes, incentives, and misinformation

14 May 2021

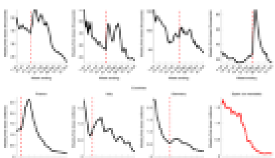
Sarah Kreps, Nabarun Dasgupta, … Douglas L. Kriner



[**Providing normative information increases intentions to accept a COVID-19 vaccine**]

09 January 2023

Alex Moehring, Avinash Collis, … Dean Eckles



[**COVID-19 vaccination mandates and vaccine uptake**]

02 June 2022

Alexander Karaivanov, Dongwoo Kim, … Hitoshi Shigeoka

- Article
- Open Access
- Published: 11 January 2023

Financial incentives for vaccination do not have negative unintended consequences

- Florian H. Schneider ,
- Pol Campos-Mercade ,
- Stephan Meier,
- Devin Pope,

- Erik Wengström &
- …
- Armando N. Meier

Show authors

Nature **volume 613**, pages 526–533 (2023)

- 8466 Accesses
- 300 Altmetric
- Metrics details

Abstract

Financial incentives to encourage healthy and prosocial behaviours often trigger initial behavioural change[1,2,3,4,5,6,7,8,9,10,11], but a large academic literature warns against using them[12,13,14,15,16]. Critics warn that financial incentives can crowd out prosocial motivations and reduce perceived safety and trust, thereby reducing healthy behaviours when no payments are offered and eroding morals more generally[17,18,19,20,21,22,23,24]. Here we report findings from a large-scale, pre-registered study in Sweden that causally measures the unintended consequences of offering financial incentives for taking the first dose of a COVID-19 vaccine. We use a unique combination of random exposure to financial incentives, population-wide administrative vaccination records and rich survey data. We find no negative consequences of financial incentives; we can reject even small negative impacts of offering financial incentives on future vaccination uptake, morals, trust and perceived safety. In a complementary study, we find that informing US residents about the existence of state incentive programmes also has no negative consequences. Our findings inform not only the academic debate on financial incentives for behaviour change but also policy-makers who consider using financial incentives to change behaviour.

Main

Offering financial incentives to encourage healthy and prosocial behaviours often triggers initial behaviour change[1,2,3,4,5,6,7,8,9,10,11], which is why financial incentives have long been considered by academics and policy-makers. For example, financial incentives have been introduced with the intent to foster blood donations[1], cancer screening rates[2], smoking cessation[3,4] and vaccination uptake[5,6,7,8,9]. However, a large and long-standing literature in the social sciences, philosophy, public health and medicine warns against offering financial incentives because of worries about a wide range of negative unintended consequences that may outweigh any initial behaviour change[12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30]. Such worries have led policy-makers and policy-advisors around the world to recommend against using financial incentives to encourage healthy and prosocial behaviours[1,31,32].

A first central concern is that financial incentives crowd out prosocial motivations, which could result in less healthy behaviour when no payments are offered and a deterioration of morals and the sense of civic responsibility more generally[12,13,14,15,16,17,25,26]. Philosopher Michael Sandel[12], for example, warns that offering financial incentives "erodes people's sense of obligation" and "diminishes the spirit of altruism". A second concern is that paying people prompts suspicion, potentially making them more hesitant to engage in certain behaviours when no payments are offered. According to this view, financial incentives signal that engaging in a health behaviour is unpleasant, risky or not as effective in improving health, and decreases trust in healthcare providers[18,19,20,21,22]. Other concerns include that incentives might change people's values[12], such as attitudes towards financial incentives, and that incentives could undermine people's sense of self-determination and make them feel coerced into a certain behaviour[23,24,28,29].

It is difficult to causally measure the unintended consequences of financial incentives. One key difficulty is finding a situation in which some people were randomly offered payments and others were not. For example, the incentive programmes that many governments introduced to increase COVID-19 vaccination uptake affected everyone at the same time and do not allow for a proper control group[33]. A second key difficulty is that studying the many potential consequences of financial incentives requires access not only to comprehensive data on people's behaviours but also to

data about individuals' morals, perceptions and feelings, which can only be measured with rich survey data.

Here we report findings from a large-scale, pre-registered study that causally measures the unintended consequences of offering financial incentives to encourage healthy and prosocial behaviour ($n = 5,019$). We overcome the identification and measurement difficulties by using a unique setting that provides random variation in exposure to incentives and by combining population-wide administrative records on health behaviours with detailed survey data. We exploit a randomized controlled trial (RCT) in the context of financial incentives for COVID-19 vaccination (P.C.-M. et al., unpublished, and ref. [5]). Participants were offered payments of 200 Swedish krona (SEK; about US $24 at the time) for taking a first dose of a COVID-19 vaccine, which increased first-dose uptake by 4 percentage points 30 days after the trial (uptake remained higher even 3 months later). The RCT setting is ideal in that it allows us to compare individuals who were randomly offered financial incentives for vaccination with individuals who were not offered any financial incentives. We combine the RCT data with new Swedish administrative records for second-dose uptake and with rich, individual-level survey data.

We document no negative impacts of offering financial incentives for taking a first dose on the timing or likelihood of participants taking the second or the third dose of a COVID-19 vaccine, for which no financial incentives were offered. We also document no effects on other health behaviours, such as blood donations and flu shots. Notably, we find no negative impacts on morals, sense of civic responsibility, trust in vaccination providers, safety and efficacy perceptions of vaccines, attitudes towards financial incentives, and feelings of self-determination and coercion. We incentivized several of the measures in the survey by implementing the choices of some participants, meaning that some of the survey measures could have real consequences and capture actual behaviour. For all outcomes, we can reject small negative impacts of 0.2 standard deviations or larger (Cohen's d), meaning that we can reject that there were even small negative consequences of offering payments for vaccination.

We complement our evidence from Sweden with evidence on the effects of large-scale incentive programmes implemented by US state governments. In

a pre-registered study in the USA ($n = 3{,}062$), participants randomly assigned to the incentives condition received detailed information about their state's COVID-19 vaccine incentive programme, whereas participants in the control condition did not receive this information. Because most of the participants were unaware that their state offered incentives for vaccination, this experimental design overcomes the identification problems by creating random variation in perceived exposure to incentives. In line with the evidence from Sweden, we find no negative impacts of being informed about incentive programmes on the willingness of participants to take a further dose, morals, trust in the state government, safety and efficacy perceptions of vaccines, or intentions to donate blood or to receive a flu shot.

The COVID-19 pandemic is the biggest health crisis in recent memory. Without very high vaccination rates, the pandemic is set to have large public health and societal impacts for years to come. With few policy tools left to motivate vaccination[34,35,36], governments, private companies and organizations across the globe have considered and introduced payments for vaccination[37]. However, evidence on whether payments for COVID-19 vaccination have negative unintended consequences, as many academics and policy-makers fear they do[17,22,24,32,37,38,39,40,41], is lacking. We report important first evidence, which is key for policy-making aimed at increasing adherence to vaccination schedules for COVID-19 vaccines, including child vaccination, recurrent booster shots for years to come, as well as for other vaccines[42].

Our findings are also important because worries about unintended consequences of payments reach well beyond vaccination[43]. Financial incentives intended to motivate healthy and prosocial behaviours have been considered in many contexts, for instance, to motivate blood[1,44,45] and organ donation[46], to curtail smoking[3,4,47], to encourage exercising and healthy eating[10,11,48], to boost medication adherence[49], to foster clinical trial participation[50,51] and to increase uptake of preventive measures, such as cancer screening[2,52]. Our findings and methods inform the large and long-standing academic literature discussing the potential negative consequences of financial incentives for behaviour change more generally.

Evidence from the Swedish RCT

Measuring unintended consequences

We use a unique combination of random variation in exposure to financial incentives from a previous RCT with comprehensive administrative and survey data (see 'Data availability' in Methods). The previous RCT was conducted from May to July 2021 and enrolled 1,131 participants who were offered SEK 200 (about US $24 at the time) to take the first dose of a COVID-19 vaccine within 30 days, forming the financial incentives condition, and 3,888 participants who were not offered any payment, forming the control condition.

We study the unintended consequences of offering financial incentives by using new administrative data collected by the Public Health Agency of Sweden on second-dose uptake and survey data on morals, safety and efficacy perceptions, feelings of self-determination and coercion, and other health behaviours. The Public Health Agency of Sweden linked the RCT data for all 5,019 participants to the COVID-19 vaccination records in late December 2021. We conducted a first survey with the RCT participants in early January 2022. Because the first survey was carried out before the participants were offered a third dose, we conducted a second survey in June 2022 on third-dose uptake. In total, 3,238 participants (2,706 participants) responded to the first survey (second survey), 726 (606) of the participants in the financial incentives condition and 2,512 (2,100) of the participants in the control condition. In both surveys, survey participation was balanced across both conditions, with no differential attrition based on personality characteristics, vaccination status, vaccine hesitancy or sociodemographics (Supplementary Information section 2.1).

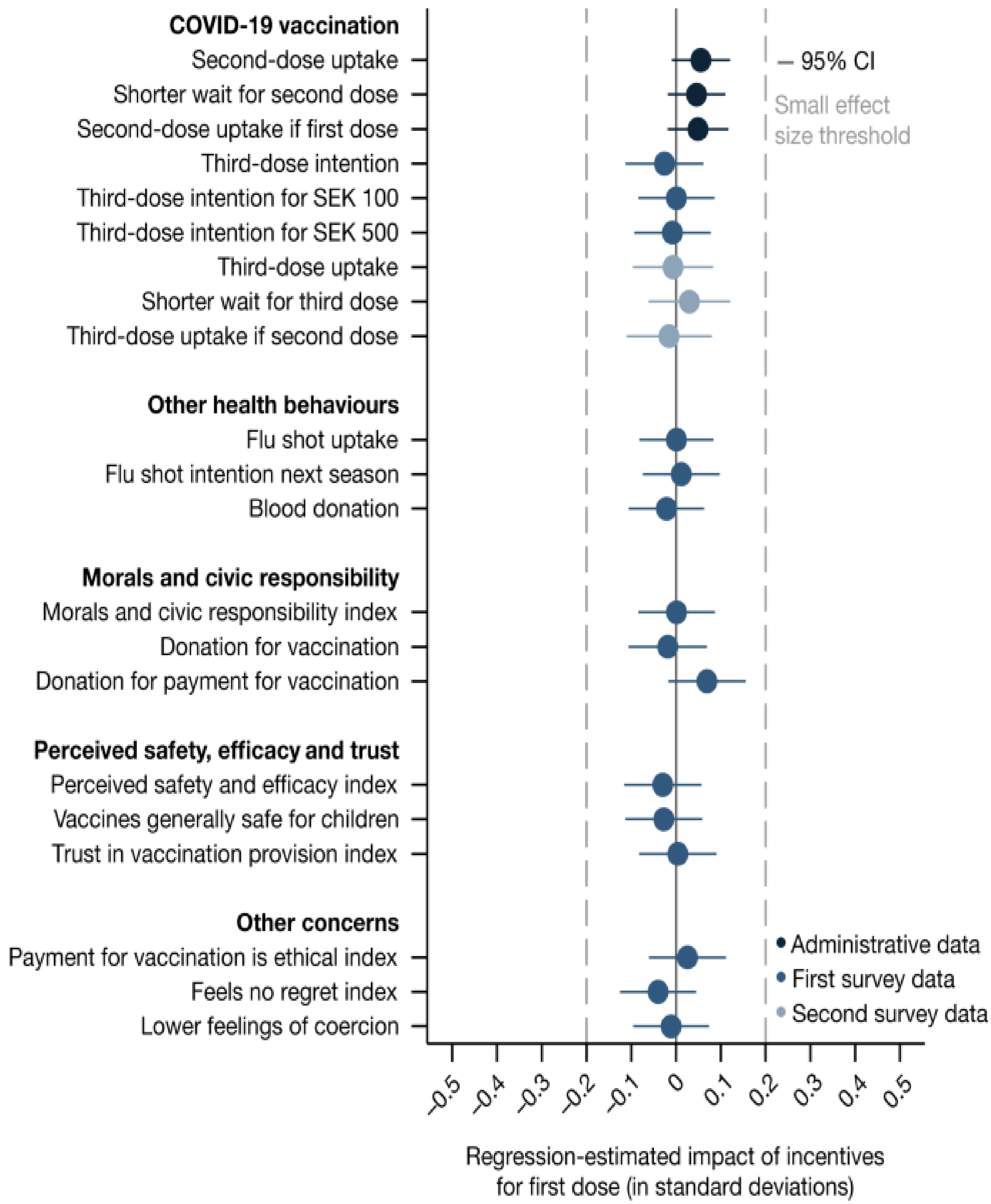
We compare the health behaviours, morals, perceptions and feelings in the financial incentives condition to the control condition. We standardize all outcomes and report all results as pre-registered. All reported results in the text, figures and tables come from ordinary least squares (OLS) regressions with heteroscedasticity-robust standard errors and all P values come from two-sided t-tests (see Methods for details). The analysis has 80% power to

detect even very small effects of −0.12 standard deviations at the 5% level, as stated in our pre-registration plan.

Results from the Swedish RCT

We first study the concern that offering financial incentives may reduce future vaccination uptake and other health behaviours when no payments are offered. Using administrative data on vaccination uptake, Fig. 1 and Table 1 show no evidence that participants in the financial incentives condition were less likely to take an unincentivized second dose of a COVID-19 vaccine (if anything, uptake increased; OLS regression, $B = 0.055$, standard error (s.e.) = 0.033, $P = 0.097$) or to delay the uptake of the second dose (OLS regression, $B = 0.046$, s.e. = 0.033, $P = 0.164$). We also do not find any effects on second-dose uptake when we restrict the sample to those who took the first dose (OLS regression, $B = 0.049$, s.e. = 0.035, $P = 0.158$). Using data from the first survey, we do not find evidence that offering monetary incentives affected the intention to take the third dose (OLS regression, $B = -0.026$, s.e. = 0.044, $P = 0.560$) nor the willingness of participants to take a third dose if they were hypothetically offered SEK 100 (OLS regression, $B = 0.001$, s.e. = 0.043, $P = 0.983$) or SEK 500 (OLS regression, $B = -0.008$, s.e. = 0.043, $P = 0.850$). Using data from the second survey, we do not find negative effects on self-reported actual third-dose uptake (OLS regression, $B = -0.007$, s.e. = 0.046, $P = 0.879$), the delay to take the third dose (OLS regression, $B = 0.030$, s.e. = 0.046, $P = 0.524$) or third-dose uptake when we restrict the sample to those who took the second dose (OLS regression, $B = -0.016$, s.e. = 0.049, $P = 0.745$).

Fig. 1: Regression-estimated effects of offering financial incentives for first-dose uptake on further COVID-19 vaccination, other health behaviours, morals and civic responsibility, perceived safety, efficacy and trust, and other concerns.



The figure is based on RCT data linked to comprehensive survey data and population-wide Swedish administrative data capturing each vaccination in Sweden. The figure shows regression-estimated effects of the financial incentives condition relative to the control condition. All regressions use the pre-registered controls consisting of gender, age, region, interactions

between age and region, being in an at-risk group for COVID-19, civil status, having children in the household, employment status, education, parents' place of birth and income (see Supplementary Information section 1.1 for details, see Supplementary Information section 2.3 and Extended Data Fig. 1 for results without controls). The blue dots indicate the estimated impact in standard deviations on the respective variables; all outcomes are defined as pre-registered. Error bars represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey lines indicate the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions across datasets are as follows: administrative data, *n* incentives = 1,132, *n* control = 3,888; first survey data, *n* incentives = 726, *n* control = 2,512; second survey data, *n* incentives = 606, *n* control = 2,100.

Source data

Table 1 Regression-estimated treatment effects of offering financial incentives for first-dose uptake, corresponding *P* values, 95% confidence intervals and equivalence tests against an effect more negative than −0.2 standard deviations

In line with these results, we find no evidence that incentives affected other health behaviours, such as flu shot uptake (OLS regression, $B = 0.001$, s.e. = 0.042, $P = 0.982$) and blood donations (OLS regression, $B = -0.021$, s.e. = 0.043, $P = 0.619$) in the previous 5 months. To summarize, we do not find that financial incentives reduced health behaviours when no payments were offered. However, these results do not address the concerns that incentives affect people more broadly, by eroding morals, decreasing safety and efficacy perceptions, and affecting feelings of self-determination and coercion.

Next, we study the concern that financial incentives could erode participants' morals and civic responsibility by using a combination of survey questions and behavioural data collected in the first survey. Figure 1 and Table 1 show no evidence that incentives affected our pre-registered index of morals and civic responsibility (OLS regression, $B = 0.001$, s.e. = 0.043, $P = 0.979$), which consists of three questions measuring participants'

sense of moral obligation to receive a COVID-19 vaccine for the good of society. We also measure altruism in the context of vaccination by offering participants the possibility to donate money to two non-governmental organizations (NGOs) that promote vaccinations. The first NGO provides COVID-19 vaccines in areas with limited access to vaccination and the second attempts to increase vaccination uptake by offering financial incentives. We do not find any differences in the amount given between the financial incentives condition and the control condition (OLS regressions, $B = -0.018$, s.e. = 0.045, $P = 0.679$ and $B = 0.069$, s.e. = 0.044, $P = 0.117$, respectively).

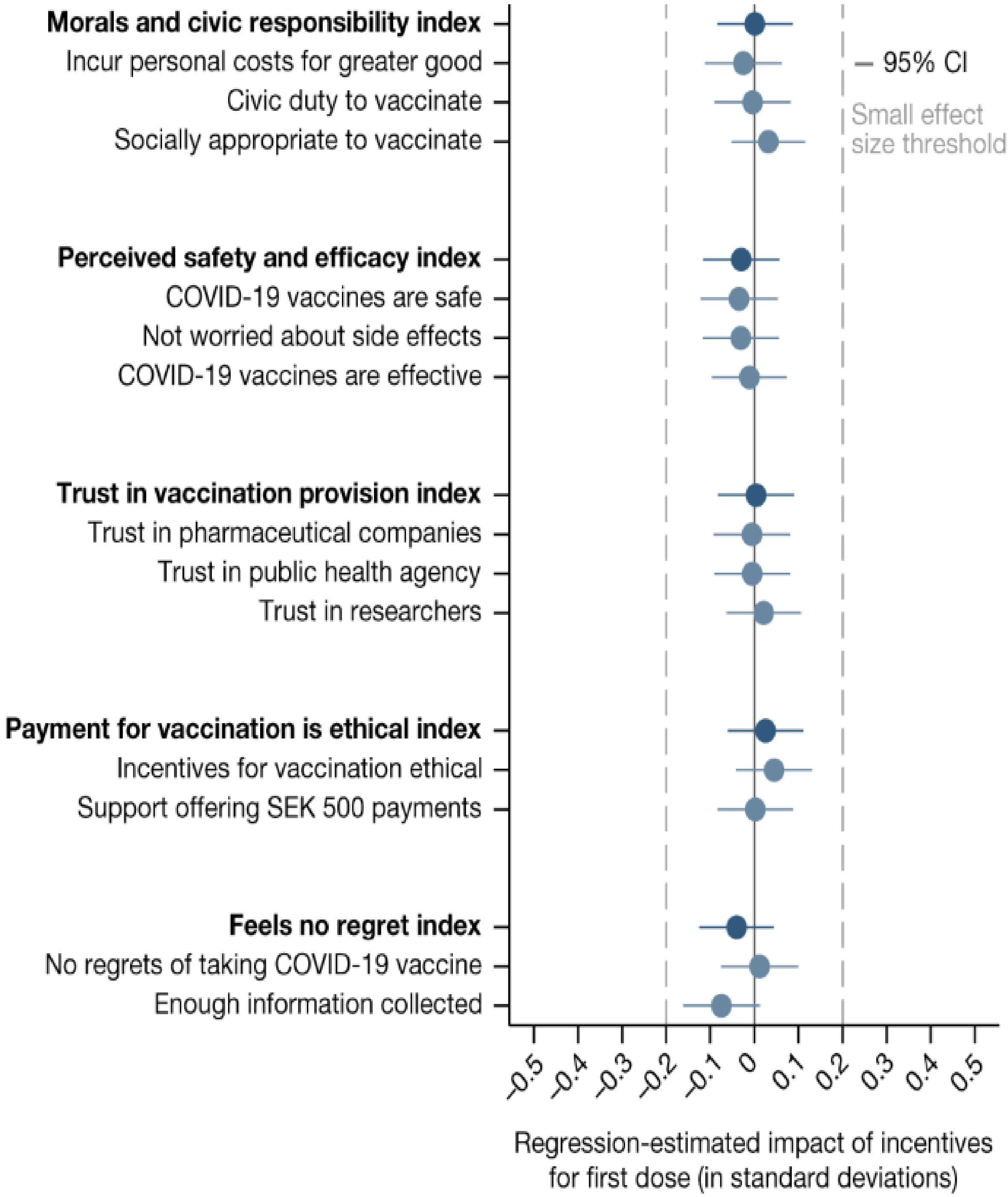
Third, we study the concern that offering financial incentives signals that vaccines are not safe and effective, which could ultimately decrease trust in vaccination providers. Figure 1 and Table 1 show no evidence that incentives affected our pre-registered index of safety and efficacy of COVID-19 vaccines (OLS regression, $B = -0.030$, s.e. = 0.044, $P = 0.499$), which includes three questions on safety perceptions, vaccine efficacy beliefs and whether participants are worried about the side-effects of COVID-19 vaccines. We also find no evidence that offering incentives affected the belief that vaccines in general are safe for children (OLS regression, $B = -0.028$, s.e. = 0.044, $P = 0.528$), nor an index capturing people's trust in researchers, the public health agency and pharmaceutical companies concerning the provision of COVID-19 vaccines (OLS regression, $B = 0.004$, s.e. = 0.044, $P = 0.928$).

Further, we study the concerns that incentives could affect participants' feelings of self-determination and coercion about their decision on whether to receive the first dose of a COVID-19 vaccine. Figure 1 and Table 1 show no evidence that participants who were offered incentives to take the first dose were more likely to say that they felt forced to take it (OLS regression, $B = -0.011$, s.e. = 0.043, $P = 0.797$) or more likely to regret their decision on whether to take the first dose (OLS regression, $B = -0.040$, s.e. = 0.043, $P = 0.352$). Finally, we also do not find any impacts on people's political views about whether paying people for vaccination is ethically acceptable (OLS regression, $B = 0.025$, s.e. = 0.044, $P = 0.560$).

Out of 21 coefficient estimates reported in Fig. 1 and Table 1, none of the 95% confidence intervals crosses the Cohen's *d* small effect size threshold of

an effect of 0.2 standard deviations. These results are robust to a battery of robustness checks, such as using each of the items underlying the pre-registered indices separately, as shown in Fig. 2, including different sets of control variables than those we pre-registered (Extended Data Fig. 1 and Supplementary Information section 2.3), and considering secondary outcome variables (Supplementary Information section 2.4).

Fig. 2: Regression-estimated effects of offering financial incentives for first-dose uptake on single items of indices.



The figure is based on RCT data linked to comprehensive survey data. The figure shows regression-estimated effects of the financial incentives condition relative to the control condition on the single items of all indices. All regressions use the pre-registered controls described in Fig. 1 and Supplementary Information section 1.1. The blue dots indicate the estimated impact in standard deviations on the respective variables. Error bars

represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey lines indicate the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions are *n* incentives = 726 and *n* control = 2,512.

Source data

Overall, Fig. 1 and Table 1 do not show any discernible negative impacts of financial incentives across the distribution of coefficient estimates. The mean coefficient is 0.004, the median coefficient is 0.001, the largest negative coefficient is −0.040 and the largest positive coefficient is 0.069, confirming the visual impression of only very small, if any, impacts. In addition, we test whether the outcomes have a different dispersion in the financial incentives condition than in the control condition and find that outcomes are not only essentially equal in means but also in distribution (for regression results and raw distributions, see Supplementary Information section 2.4.1, Supplementary Figs. S12–S15 and section 2.4.2).

Equivalence testing further demonstrates that there were no meaningful negative impacts across all outcomes. For the equivalence testing[53,54], we use the standard effect size threshold for small effects of 0.2 standard deviations as the smallest effect size of interest. Table 1 shows that tests for all outcomes are highly statistically significant, clearly rejecting negative impacts more negative than −0.2 standard deviations (largest $P = 0.0001$). The test results are similar when we specify the smallest effect size of interest as the smallest effect size that our study design can reliably detect based on the pre-registration[54] (see the Supplementary Information section 2.2.4 for details). Overall, the results provide strong evidence against even small negative consequences of offering payments for vaccination.

No impacts on different groups

We further explore whether any of the treatment effects differ based on variables measured before offering incentives, including vaccine hesitancy, as well as sociodemographics such as income, education, age and gender. The theoretical concerns in the literature on unintended consequences concern mainly individuals with positive vaccination attitudes; these are the

individuals whose prosocial and intrinsic motivation could be crowded out and who might start doubting the safety and efficacy of vaccines. On the other hand, offering incentives might make the hesitant even more sceptical of vaccination.

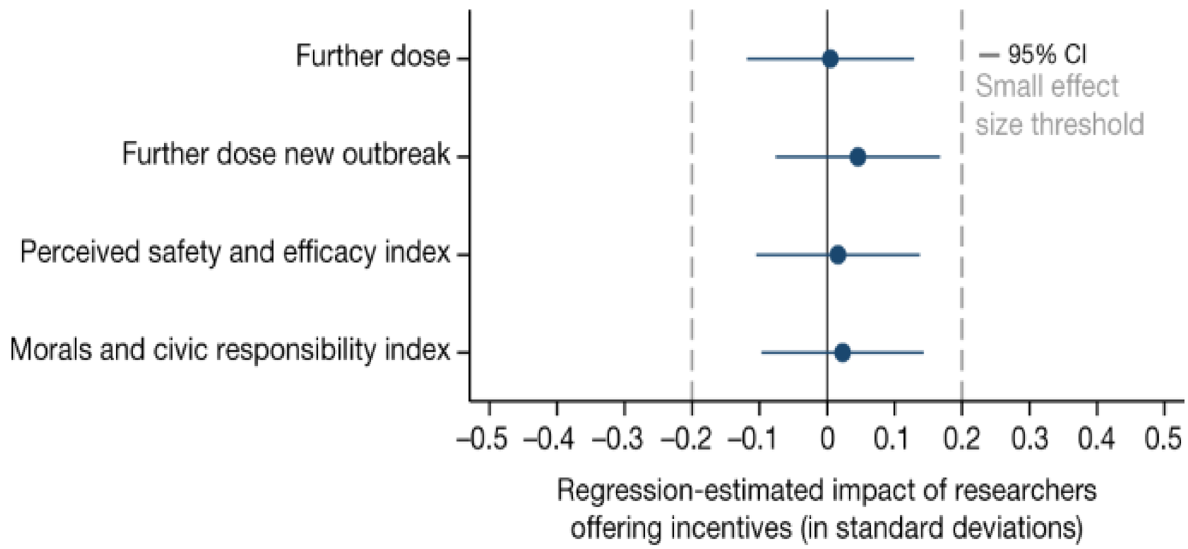
The data do not indicate consistent negative effects for either relatively vaccine positive or hesitant groups. A potential limitation of these findings is that we do not study a very hesitant population[55,56]. In addition, we find similarly muted treatment effects across sociodemographic subgroups. Overall, we do not find that any of the groups suffered from negative unintended consequences (Extended Data Tables 1–4; for further details and regression results, see Supplementary Information section 2.4.3).

Different entities offering incentives

An open question is whether the impact of incentives differs when paid by public institutions rather than researchers. We examine this question in a complementary study in Sweden (n = 1,001). We use the fact that the previous RCT was implemented in a collaboration between researchers and the Public Health Agency of Sweden. Some study participants were informed that “a team of researchers participated in the implementation of the incentive programme”, whereas others were told that the Public Health Agency of Sweden did so (see Supplementary Information section 1.1.6 for details).

As shown in Fig. 3, we find no evidence that people’s reactions to financial incentives depended on whether they were informed that the public health authorities or researchers offered the payments. Equivalence testing further shows that we can clearly reject even small negative treatment effects of 0.2 standard deviations (Supplementary Information section 2.5).

Fig. 3: Regression-estimated effects of informing Swedish residents about researchers versus the public health authorities being involved in offering vaccination incentives on further COVID-19 vaccination, morals and civic responsibility, and perceived safety and efficacy.



The figure is based on experimental data from a general population sample of Swedish residents. The figure shows regression-estimated effects of the researcher condition (informing participants that researchers participated in the implementation of an incentive programme) relative to the government condition (informing participants that the Public Health Agency of Sweden participated in the implementation of an incentive programme), as pre-registered. All regressions use controls consisting of gender, age, education and income (see Supplementary Information section 2.5 for results without controls). The blue dots indicate the estimated impact in standard deviations on the respective variables; all outcomes are defined as pre-registered. Error bars represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey lines indicate the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions are *n* researcher = 515 and *n* government = 486.

Source data

Evidence from US incentive programmes

Informing people about incentive programmes

In 2021, many US states introduced financial incentives, ranging from small, guaranteed rewards to lotteries that gave vaccinated individuals a chance to

win large prizes[57]. Much of the debate in the USA about unintended consequences of financial incentives has focused on these state incentive programmes[22,24,37,41]. In this section, we complement the evidence from Sweden by studying whether learning about the existence of US state incentive programmes had unintended consequences.

Worries about the unintended consequences of monetary incentives apply to this setting as well. Learning about the existence of a state incentive programme could, for example, signal to participants that selfishness is an appropriate response—thereby eroding morals—or that being vaccinated is risky. This could in turn reduce future vaccination uptake. Moreover, although there is debate about the success of some of these state incentive programmes, with mixed empirical evidence[57,58,59,60,61,62], unintended consequences can occur in either case (see the discussion in Supplementary Information section 2.6.9). For instance, individuals may not be more likely to vaccinate in response to incentives but may grow more suspicious of vaccinations or change their vaccination morals.

We conducted a pre-registered experiment in June and July 2022 using a general population sample from 12 states that implemented vaccine incentive programmes ($n = 3{,}062$). We use the fact that many people in the USA (62.3% in our sample) were unaware that state governments rolled out financial incentive programmes, as states often did not publicize the programmes aggressively[58].

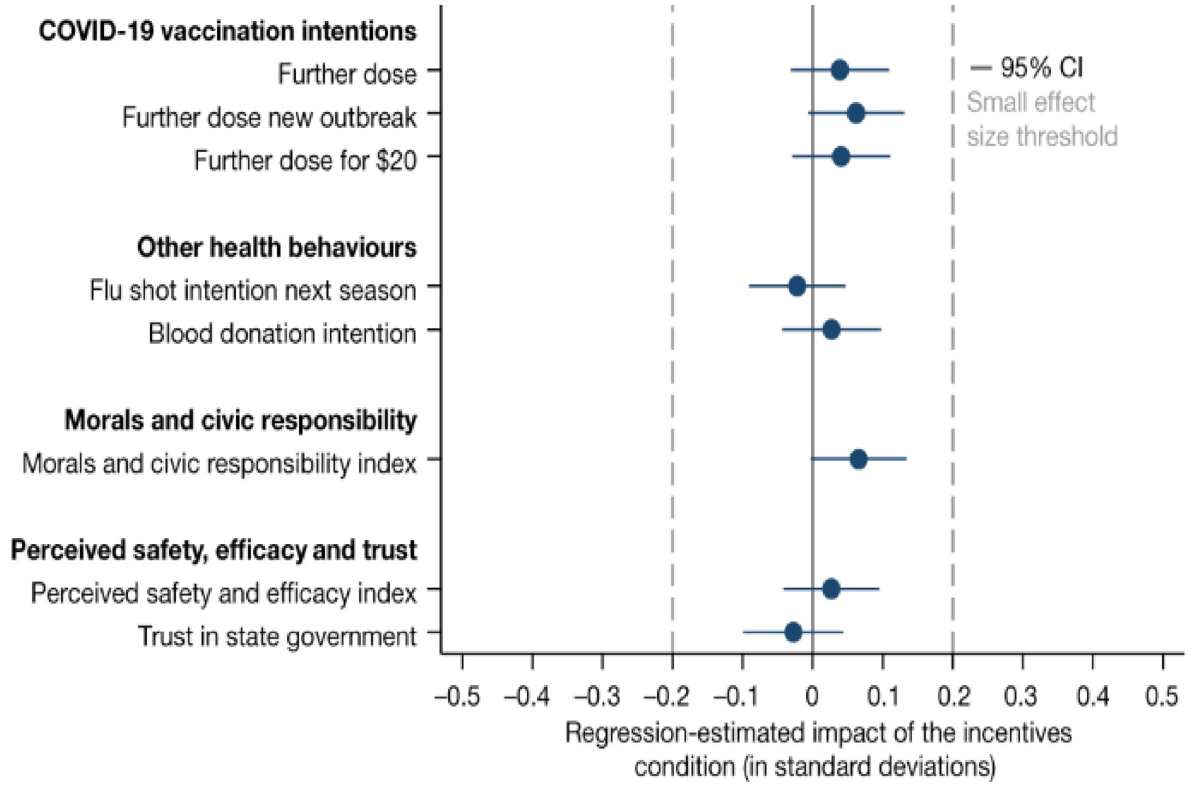
We randomly allocated participants to two treatment conditions, the incentives condition and the control condition (see Methods for details). The participants randomly assigned to the incentives condition received detailed information about their state's COVID-19 vaccine incentive programme, whereas participants in the control condition did not receive this information. For example, treated participants who resided in California in 2021 were told that the government of California implemented the 'Vax for the Win' programme, which distributed more than $100 million in cash prizes and $50 gift or grocery cards from May 2021 to January 2022. Such provision of information creates random variation in perceived exposure to incentives.

We analysed whether the provision of information had any impact on the willingness of participants to receive future shots of a COVID-19 vaccine, morals and safety perceptions. To avoid experimenter demand effects, we measured these outcomes in an apparently unrelated follow-up survey about 5 days after the survey in which we provided the information (Methods). The data from the follow-up survey shows that participants who received information about the state incentive programmes were still aware of them 5 days later (Extended Data Fig. 2 and Supplementary Information section 2.6.3). All reported results in the text, figures and tables come from OLS regressions with heteroscedasticity-robust standard errors and all P values come from two-sided t-tests (see Methods for details).

Results from the US study

Figure 4 and Table 2 show no evidence that participants in the incentives condition who were informed about the existence of US state incentive programmes were less willing to receive a further dose within the next 6 months (OLS regression, $B = 0.039$, s.e. = 0.036, $P = 0.276$), to receive a further dose in case there would be a new outbreak (if anything, their willingness increased; OLS regression, $B = 0.062$, s.e. = 0.035, $P = 0.077$) or to receive a further dose if their state government offered them $20 for it (OLS regression, $B = 0.041$, s.e. = 0.036, $P = 0.256$). We also do not find that participants in the incentives condition were less willing to take a flu shot next winter (OLS regression, $B = -0.022$, s.e. = 0.035, $P = 0.528$) or to donate blood (OLS regression, $B = 0.027$, s.e. = 0.036, $P = 0.454$). Finally, we find no evidence that incentives eroded participants' morals and civic responsibility (if anything, morals improved; OLS regression, $B = 0.066$, s.e. = 0.035, $P = 0.060$), their safety and efficacy perceptions about the COVID-19 vaccines (OLS regression, $B = 0.027$, s.e. = 0.035, $P = 0.448$) or their trust in the state government (OLS regression, $B = -0.027$, s.e. = 0.036, $P = 0.450$).

Fig. 4: Regression-estimated effects of informing US residents about state vaccination incentive programmes on further COVID-19 vaccination, other health behaviours, morals and civic responsibility, and perceived safety, efficacy and trust.



The figure is based on experimental data from a general population sample of US residents in 12 states that introduced incentive programmes for COVID-19 vaccination. The figure shows regression-estimated effects of the incentives condition (informing participants about the existence of incentive programmes in their state) relative to the control condition, as pre-registered. All regressions use controls consisting of gender, age, education, employment status, income and state of residence in 2021 (see Extended Data Fig. 3 for results without controls). The blue dots indicate the estimated impact in standard deviations on the respective variables; all outcomes are defined as pre-registered. Error bars represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey lines indicate the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions are *n* incentives = 1,521 and *n* control = 1,541.

Source data

Table 2 Regression-estimated effects of informing US residents about state vaccination incentive programmes, corresponding *P* values, 95%

confidence intervals and equivalence tests against an effect more negative than −0.2 standard deviations

Overall, Fig. 4 and Table 2 do not show any discernible negative impacts of receiving information about the existence of US state incentive programmes. In Supplementary Information section 2.6, we show that our results are robust to a battery of robustness checks, such as including different sets of control variables (Extended Data Fig. 3), using each of the items underlying the pre-registered indices separately (Extended Data Fig. 4) and using different inclusion criteria.

Out of all coefficient estimates reported in Fig. 4 and Table 2, none of the 95% confidence intervals crosses the Cohen's *d* small effect size threshold of an effect of 0.2 standard deviations. Equivalence testing confirms strong evidence for the absence of meaningful treatment effects across all outcomes (see Table 2 and Supplementary Information section 2.6.5 for details).

Finally, we explore whether any of the treatment effects differ based on vaccine hesitancy, political attitudes and sociodemographics. We also study potential heterogeneous impacts based on participants' state of residence, which allows us to examine whether there were unintended consequences in states in which incentive programmes were more or less successful. We find that treatment effects are similarly mute across all subgroups and all states (Supplementary Information sections 2.6.8 and 2.6.9).

Discussion and conclusions

Our studies document that offering modest payments for vaccination has no sizable unintended consequences. We can reject even small negative impacts of financial incentives for COVID-19 vaccination on people's future vaccination uptake, other health behaviours, morals and civic responsibility, perceived safety and effectiveness of the vaccines, trust in vaccine providers, and feelings of self-determination and coercion.

Many healthy and prosocial behaviours, such as donating blood, not smoking and vaccinating, have large individual and societal consequences[1,3,5,6,63,64]. Offering financial incentives for behaving healthily

and prosocially is widely considered by policy-makers to change behaviour[1,5,31,32,37]. Although offering financial incentives often triggers initial behaviour change, a large academic literature warns against using financial incentives because of unintended consequences. This tension puts policy-makers in a tough spot over the extent to which they should heed or ignore the warnings when considering introducing financial incentives to encourage behaviour change. During the COVID-19 pandemic, for instance, many governments and organizations worldwide offer financial incentives for vaccination, whereas others abstain, worried about grave unintended consequences[37]. Our study provides important evidence that will allow policy-makers to make more informed decisions when weighing the costs and benefits of introducing financial incentives to change behaviour.

Although much of the academic discussion focuses on the negative unintended consequences, financial incentives could, in principle, also have positive unintended consequences[45,48,65]; incentives might not only trigger initial behavioural change but could positively affect future health behaviours, morals, perceptions and feelings. However, we find no support for positive unintended consequences of financial incentives for COVID-19 vaccination.

The evidence from the Swedish RCT and the US state incentive programmes complement each other by using samples with different characteristics, applying different methodologies and studying incentive programmes that differ in scale, incentive type and entity that offers the incentives. Our findings that financial incentives for COVID-19 vaccination do not have negative unintended consequences in both contexts, as well as the lack of consistent negative treatment effects across different sociodemographic and vaccine hesitancy groups, speak to the generalizability of our findings.

However, several limitations remain. First, our studies rely on samples from high-income Western countries. The results may not generalize to low-income countries or to countries with meagre social security systems. Second, to collect encompassing survey data as well as to guarantee compliance with ethical and consent guidelines, participants in our studies were aware that they participated in a study. This awareness could, in principle, affect results but it can hardly be avoided when linking survey with administrative records. Third, our paper focuses on financial incentives

for COVID-19 vaccination. Although this is a particularly relevant context given the current debate, we hope that the paper motivates new studies across different contexts (such as organ donation or cancer screening) to improve our understanding of the consequences of offering incentives. Last, although our evidence also informs the normative debate of whether paying for vaccination is ethically permissible[24,25], ethical debates will not be resolved by empirics alone[66].

Despite its limitations, our study has a clear finding: offering modest financial incentives for vaccination has limited, if any, negative unintended consequences. Contrary to prominent warnings in the academic literature and public debate, our work suggests that modest financial incentives for vaccination can be used without worries about grave unintended consequences.

Methods

Swedish RCT approval and pre-registration

We conducted a pre-registered study with a general population sample of Swedish residents. In two online surveys, we recruited participants from a sample that took part in an earlier RCT[5]. In this earlier RCT, participants were randomly allocated to either a financial incentives condition that offered payments of 200 SEK (about US $24 at the time) conditional on receiving the first dose of a COVID-19 vaccine or a control condition that did not offer any financial incentives (see Supplementary Information section 1.1.5 for details). This earlier RCT provides us with random assignment of participants to financial incentives for taking a first dose. We match the RCT data with exhaustive population-wide Swedish administrative records of COVID-19 vaccinations, which allow us to examine whether and when each of the participants received an unincentivized second dose of a COVID-19 vaccine. We then matched these data with data from two online surveys, in which we measured participants' health behaviours, morals, perceptions and feelings.

The Swedish ethical review authority (Etikprövningsmyndigheten) approved the protocols of the study (reference number 2021-06367-02). Participants

were informed that the study was conducted by researchers and that their data would be matched with vaccination registries by the public health authorities. Informed consent was obtained from all study participants as part of the survey.

We pre-registered the data collection and analysis at the AEA RCT Registry (http://www.socialscienceregistry.org/trials/8727 and http://www.socialscienceregistry.org/trials/9580). Our analysis closely follows the pre-registration plan. In the main analysis, we use the following pre-registered linear regression to estimate treatment effects:

$${Y}_{i}={\beta }_{0}+{\beta }_{1}\times {I}_{i}+{X}_{i}\gamma \text{'}+{{\epsilon }}_{i}$$

in which Y_i captures the outcome variable for participant i, I_i is a dummy capturing whether participant i is in the financial incentives condition, β_1 estimates the effect of incentives on the outcome variable and ε_i is an individual-specific error. The vector X_i is the vector of pre-registered controls to reduce variability, consisting of participant i's gender, age, region, interactions between age and region, being in an at-risk group for COVID-19, civil status, having children in the household, employment status, education, parents' place of birth and income (see Supplementary Information section 1.1 for definitions of all variables and further details about the data analysis). We estimate treatment effects using OLS regressions with heteroscedasticity-robust standard errors.

In Supplementary Information section 2.3, we show that the results are robust to including no controls (Extended Data Fig. 1), different sets of control variables, using sample weights and using different inclusion criteria. The battery of further analyses shows no negative unintended consequences of offering financial incentives for vaccination.

Administrative vaccination records

We use administrative data from COVID-19 national vaccination registers comprising all residents of Sweden. The administrative records include the date of each COVID-19 vaccination of each resident. As it is not possible to

opt out of or delete records in the vaccination registry, the administrative records include whether and when each participant received the second dose of a COVID-19 vaccine. Notably, the participants were not offered any payments to take the second dose. Note also that individuals in Sweden had to book the appointments for the first and second doses separately (see Supplementary Information section 1.1.4 for details). The Public Health Agency of Sweden linked the previous RCT data at the individual level with the administrative data on 21 December 2021. As the previous RCT ended on 13 July, we observe for participants whether and when they received the second dose of a COVID-19 vaccine within a time window of 158 days after participation in the trial.

We constructed the following outcomes based on administrative data:

- Second-dose uptake: we measured whether participants took the second dose of a COVID-19 vaccine after participation in the RCT.
- Shorter wait for second dose: we measured how long the participants waited until they received the second dose. For participants who did not take a second dose, we used the maximum wait time that we could observe. We then reverse-coded the outcome so that a positive coefficient indicates shorter wait time.
- Second-dose uptake if first dose: this outcome corresponds to second-dose uptake but we restricted the sample to participants who took a first dose ($n = 4,358$).

We pre-registered second-dose uptake as a main outcome measure based on administrative data.

Surveys

The survey participants were recruited from a general population panel in Sweden by the survey company Norstat. Norstat actively recruits people by means of phone calls to create a representative panel in terms of age, region and gender. For both surveys, we asked the company to recruit as many participants as possible from the incentives ($n = 1,131$) and control conditions ($n = 3,888$) from the previous RCT (the control condition

includes the control and no-reminders conditions from the RCT; see Supplementary Information section 1.1.2 for details). The surveys were programmed in Qualtrics. We provide the questionnaires translated into English in Supplementary Information section 3.1.

Participants in the first survey were paid SEK 50 (about $5.5) for a 10-min survey. Responses were collected in early January 2022. In total, 726 of the participants in the financial incentives condition and 2,512 of the participants in the control condition responded to the survey. Participants in the second survey were paid SEK 10 (about $1) for a 2-min survey. Responses were collected in late June 2022. In total, 606 of the participants in the financial incentives condition and 2,100 of the participants in the control condition responded to the second survey. In both surveys, survey participation was balanced across both conditions, with no differential attrition based on personality characteristics, vaccination status, vaccine hesitancy or sociodemographics (Supplementary Information section 2.1). The survey completion rates were greater than 99% for each survey, with no differences across the incentives and control conditions (Supplementary Information section 2.1).

The participants from the previous RCT are on average 34.6 years old, have an average monthly income of SEK 24,724 and consist of 42% men. In comparison with the Swedish population (in the desired age range of 18–49 years), our sample is representative with respect to age, income and region. However, we have a slight overrepresentation of women and people with a college education and an underrepresentation of people with immigrant background (Supplementary Information section 2.1). In Supplementary Information section 2.3, we show that the results do not change when using sampling weights to adjust for sample composition. In addition, we find that the sociodemographics of participants are comparable across experimental conditions (Supplementary Information section 2.1).

In the first survey, we measured participants' behaviours, morals, perceptions and feelings related to COVID-19 vaccination. For some outcomes, we aggregated several items into an index, exactly as pre-registered. We measured the following main survey outcomes:

- Third-dose intention: we asked participants whether they are planning to take the third dose of a COVID-19 vaccine (booster shot) when it becomes available to them.

- Third-dose intention for SEK 100/SEK 500: we asked participants to assume that their region pays SEK 100/SEK 500 for everyone who takes the third dose and asked them how likely they would be to take the third dose.

- Flu shot uptake: we asked participants whether they have taken a flu shot in the past 5 months.

- Flu shot intention next season: we asked participants how likely they are to receive a flu shot next season (from fall 2022 to spring 2023).

- Blood donation: we asked participants whether they donated blood in the past 5 months.

- Morals and civic responsibility index: we aggregated the answers to the following items on morals and civic responsibility: (i) I am willing to take the personal costs of receiving a COVID-19 vaccine (such as time, discomfort, mild side effects) for the greater good of society; (ii) I think people have a civic duty or a moral obligation to receive a COVID-19 vaccine; (iii) not taking a COVID-19 vaccine would be generally viewed as socially inappropriate.

- Perceived safety and efficacy index: we aggregated the answers to the following three risk and efficacy perceptions: (i) in general, COVID-19 vaccines are safe; (ii) I am worried about the side effects from COVID-19 vaccines (reverse-coded); (iii) COVID-19 vaccines are highly effective at protecting my health.

- Vaccines generally safe for children: we asked participants whether they think, in general, vaccines given to children, such as the measles vaccine, are safe for healthy children.

- Trust in vaccination provision index: we aggregated the answers to the following three questions on trust: when it comes to the COVID-19

vaccine process, I trust: (i) the pharmaceutical or drug companies; (ii) the researchers studying the effects of the vaccines; (iii) the Public Health Agency of Sweden.

- Feels no regret index: we aggregated the answers to the following two items: (i) we asked participants whether they regret the decision they made on whether to receive the first dose of a COVID-19 vaccine (reverse-coded); (ii) we asked participants whether they gathered enough information to feel well informed about the benefits and risks of the vaccine when deciding to receive the first dose of a COVID-19 vaccine or not. The first question is based on a survey item taken from Ambuehl et al.[67].

- Lower feelings of coercion: we asked participants whether, when deciding to receive the first dose of a COVID-19 vaccine or not, they felt forced to take or not take the COVID-19 vaccine (reverse-coded). This question is based on a survey item taken from Ambuehl et al.[67].

To aggregate the individual items into the indices, we standardized each item (subtracted the mean and then divided it by the standard deviation), added the items and divided the result by the number of items. We then standardized all outcomes, including the indices, such that effect sizes are comparable across outcomes.

We also measured the following behaviours, which we standardized for the analysis:

- Donation for vaccination: subjects divided SEK 100 between themselves and the Global Alliance for Vaccines and Immunization. The Global Alliance for Vaccines and Immunization collects donations to provide COVID-19 vaccines in areas with otherwise limited access to vaccination. We incentivized this question by implementing the choice of ten randomly drawn participants.

- Donation for payment for vaccination: subjects divided SEK 100 between themselves and the New Incentives organization. The New Incentives organization is a NGO that attempts to increase vaccination uptake for diseases such as measles by paying people for being

vaccinated. We incentivized this question by implementing the choice of ten randomly drawn participants.

- Payment for vaccination is ethical index: we aggregated the answers to the following two items: (i) financial rewards for vaccinating against COVID-19 are unethical (reverse-coded); (ii) I would support the introduction of monetary payments of SEK 500 for those who get vaccinated (or are already vaccinated) against COVID-19. We followed the approach by Elías et al.[68] and told participants that their views would be shared with policy-makers.

We pre-registered third-dose intention, perceived safety and efficacy index, and morals and civic responsibility index as the main survey outcome variables of this survey.

In the second survey, we measured participants' third-dose vaccination uptake. We measured the following survey outcomes:

- Third-dose uptake: we asked participants whether they took the third dose of a COVID-19 vaccine.

- Shorter wait for third dose: we asked participants when they took the third dose of a COVID-19 vaccine.

- Third-dose uptake if second dose: this outcome corresponds to "third-dose uptake" but we restricted the sample to participants who took a second dose ($n = 2{,}463$).

Supplementary Information section 2.4.1 gives the distributions of all survey measures.

Swedish complementary study

In June 2022, we conducted a pre-registered online study using a general population sample of 1,001 Swedish participants (similar to the sample of the previous RCT; 46% men, average age = 31.56 years, standard deviation = 8.47) recruited by the survey company Norstat. The study examined whether people react differently when they are told that the government or

researchers paid people for COVID-19 vaccination. We use the fact that most people in Sweden are unaware of the previous Swedish RCT and that the previous RCT was implemented in collaboration with a governmental organization.

The study randomly allocated participants into two treatment conditions, the government condition and the researcher condition. In both conditions, we first described the earlier RCT. The participants in the researcher condition were then told that “a team of researchers participated in the implementation of the incentive programme”, whereas the participants in the government condition were told that “the Public Health Agency of Sweden participated in the implementation of the incentive programme”. Finally, we measured our outcome measures, represented in Fig. 3. See Supplementary Information section 1.1.6 for a more detailed description of the study.

We pre-registered the data collection and analysis at the AEA RCT Registry (http://www.socialscienceregistry.org/trials/9584). Our analysis closely followed the pre-registration plan. Our analysis has 80% power to detect smaller effects than 0.2 standard deviations at the 5% level, as stated in our pre-registration plan. The Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich approved the protocols of the complementary study (reference number 2022-045). Informed consent was obtained from all study participants as part of the survey.

US state incentive programmes study

In June and July 2022, we conducted a pre-registered study with 3,062 participants from a general population sample of US residents to study whether COVID-19 financial incentive programmes implemented by US states had negative unintended consequences. We use the fact that many people in the USA (around 62.3% in our sample) are unaware that state governments implemented financial incentive programmes. The survey was programmed in Qualtrics. We provide the questionnaire items in Supplementary Information section 3.

We recruited participants from 12 states that implemented vaccine incentive programmes either at the state or the county level: California, Florida,

Illinois, Kentucky, Louisiana, Michigan, Missouri, New York, North Carolina, Ohio, Pennsylvania and Texas (see Supplementary Information sections 1.2.2 and 3.3 for a description of the state incentive programmes).

In the study, we first measured participants' sociodemographics, including their state of residence in 2021. Next, we measured whether participants knew about the existence of state incentive programmes; we asked them whether, in 2021, any governmental organization in their state offered any financial compensation to people who were vaccinated against COVID-19. We continued by eliciting COVID-19 vaccination history and vaccination attitudes. Finally, we randomly allocated participants into one of two treatment conditions: the incentives condition or the control condition. The participants in the incentives condition received detailed information about their state government's COVID-19 vaccine incentive programme (Supplementary Information section 3), whereas participants in the control condition did not receive this information. This procedure creates random variation in perceived exposure to incentives, allowing us to study the unintended consequences of being exposed to incentives (for treatment effects on awareness, see Extended Data Fig. 2).

To avoid experimenter demand effects[69], we elicited the outcome measures in an ostensibly unrelated second study[70]. We blur the connection between the two surveys by letting 4–6 days pass between the two surveys and by using different fonts, formats and university affiliations. In the second survey, we elicited our outcome measures. We measured flu shot intention next season and all survey items included in the morals and civic responsibility index and the perceived safety and efficacy index. In addition, we elicited the following measures:

- Further dose: we asked participants whether they planned to take a further COVID-19 vaccine dose (regardless of the number of doses they received in the past) within the next 6 months.

- Further dose new outbreak: we told participants to assume that there would be a new outbreak of the COVID-19 pandemic in 6 months and the Centers for Disease Control and Prevention would recommend people to take a further COVID-19 vaccine dose (regardless of the

number of doses they received in the past). We asked participants whether, in this situation, they would take a further dose.

- Further dose for $20: we told participants to assume that there would be a new outbreak of the COVID-19 pandemic in 6 months, the Centers for Disease Control and Prevention would recommend people to take a further COVID-19 vaccine dose (regardless of the number of doses they received in the past) and that every person receiving a further dose would receive $20. We asked participants whether, in this situation, they would take a further dose.

- Blood donation intention: we asked participants whether they plan to donate blood in the next 6 months.

- Trust in state government: we asked participants how much trust they have in the government of their state of residence when it comes to handling problems.

The survey participants were recruited from a general population panel in the USA by the survey company Prolific. Participants in the first survey were paid $1 for a 4-min survey and participants in the follow-up survey were paid $0.5 for a 2-min survey. In total, 3,980 people responded to the first survey and 3,062 people responded to the follow-up survey. We can therefore match the two surveys for 3,062 participants (50% men, average age = 36.76 years, standard deviation = 13.54, 41% Democrats). Participation in the follow-up survey was balanced across both conditions, with no differential attrition based on vaccination status, vaccine hesitancy or sociodemographics (Supplementary Information section 2.6.1).

We pre-registered the data collection and analysis at the AEA RCT Registry (http://www.socialscienceregistry.org/trials/9607). Our analysis closely follows the pre-registration plan. We use a linear regression to estimate treatment effects using OLS regressions with heteroscedasticity-robust standard errors. We control for gender, age, education, employment status, income and state of residence in 2021. In Supplementary Information section 2.6, we show that the results are robust to including no controls and different sets of control variables. Our analysis has 80% power to detect

smaller effects than 0.2 standard deviations at the 5% level, as stated in our pre-registration plan.

The Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich approved the protocols of the study (reference number 2022-045). Participants were informed that the study was conducted by researchers and informed consent was obtained from all study participants as part of the survey.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in the analyses and figures in the article are available on Zenodo at https://doi.org/10.5281/zenodo.7214856. Source data are provided with this paper.

Code availability

The code to replicate the analyses and figures in the article is available on Zenodo at https://doi.org/10.5281/zenodo.7214856. Analyses were conducted using STATA 16.

References

1. Lacetera, N., Macis, M. & Slonim, R. Economic rewards to motivate blood donations. *Science* **340**, 927–928 (2013).

2. Stone, E. G. et al. Interventions that increase use of adult immunization and cancer screening services: a meta-analysis. *Ann. Intern. Med.* **136**, 641–651 (2002).

3. Volpp, K. G. et al. A randomized, controlled trial of financial incentives for smoking cessation. *N. Engl. J. Med.* **360**, 699–709 (2009).

4. Halpern, S. D. et al. A pragmatic trial of e-cigarettes, incentives, and drugs for smoking cessation. *N. Engl. J. Med.* **378**, 2302–2310 (2018).

5. Campos-Mercade, P. et al. Monetary incentives increase COVID-19 vaccinations. *Science* **374**, 879–882 (2021).

6. Banerjee, A. V., Duflo, E., Glennerster, R. & Kothari, D. Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *Br. Med. J.* **340**, c2220 (2010).

7. Alsan, M., Garrick, O. & Graziani, G. Does diversity matter for health? Experimental evidence from Oakland. *Am. Econ. Rev.* **109**, 4071–4111 (2019).

8. Klüver, H., Hartmann, F., Humphreys, M., Geissler, F. & Giesecke, J. Incentives can spur COVID-19 vaccination uptake. *Proc. Natl Acad. Sci.* **118**, e2109543118 (2021).

9. Serra-Garcia, M. & Szech, N. Incentives and defaults can increase COVID-19 vaccine intentions and test demand. *Manage. Sci.* https://doi.org/10.1287/mnsc.2022.4405 (2022).

10. Volpp, K. G. et al. Financial incentive-based approaches for weight loss: a randomized trial. *JAMA* **300**, 2631–2637 (2008).

11. Charness, G. & Gneezy, U. Incentives to exercise. *Econometrica* **77**, 909–931 (2009).

12. Sandel, M. *What Money Can't Buy: The Moral Limits of Markets* (Macmillan, 2012).

13. Titmuss, R. M. *The Gift Relationship: From Human Blood to Social Policy* (Policy Press, 1970).

14. Satz, D. *Why Some Things Should Not Be for Sale: The Moral Limits of Markets* (Oxford Univ. Press, 2010).

15. Bowles, S. Policies designed for self-interested citizens may undermine “the moral sentiments”: evidence from economic experiments. *Science* **320**, 1605–1609 (2008).

16. Frey, B. S. & Jegen, R. Motivation crowding theory. *J. Econ. Surv.* **15**, 589–611 (2001).

17. Largent, E. A. & Miller, F. G. Problems with paying people to be vaccinated against COVID-19. *JAMA* **325**, 534–535 (2021).

18. Frey, B. S. & Oberholzer-Gee, F. The cost of price incentives: an empirical analysis of motivation crowding-out. *Am. Econ. Rev.* **87**, 746–755 (1997).

19. Bénabou, R. & Tirole, J. Intrinsic and extrinsic motivation. *Rev. Econ. Stud.* **70**, 489–520 (2003).

20. Ellingsen, T. & Johannesson, M. Pride and prejudice: the human side of incentive theory. *Am. Econ. Rev.* **98**, 990–1008 (2008).

21. Cryder, C. E., London, A. J., Volpp, K. G. & Loewenstein, G. Informative inducement: study payment as a signal of risk. *Soc. Sci. Med.* **70**, 455–464 (2010).

22. Volpp, K. G. & Cannuscio, C. C. Incentives for immunity—strategies for increasing COVID-19 vaccine uptake. *N. Engl. J. Med.* **385**, e1 (2021).

23. Deci, E. L. Effects of externally mediated rewards on intrinsic motivation. *J. Pers. Soc. Psychol.* **18**, 105–115 (1971).

24. Jecker, N. S. Cash incentives, ethics, and COVID-19 vaccination. *Science* **374**, 819–820 (2021).

25. Savulescu, J. Good reasons to vaccinate: mandatory or payment for risk? *J. Med. Ethics* **47**, 78–85 (2021).

26. Fehr, E. & Rockenbach, B. Detrimental effects of sanctions on human altruism. *Nature* **422**, 137–140 (2003).

27. Böhm, R. et al. Crowdsourcing interventions to promote uptake of COVID-19 booster vaccines. *EClinicalMedicine* **53**, 101632 (2022).

28. Falk, A. & Kosfeld, M. The hidden costs of control. *Am. Econ. Rev.* **96**, 1611–1630 (2006).

29. Roth, A. E. Repugnance as a constraint on markets. *J. Econ. Perspect.* **21**, 37–58 (2007).

30. Gneezy, U. & Rustichini, A. Pay enough or don't pay at all. *Q. J. Econ.* **115**, 791–810 (2000).

31. World Health Organization (WHO). *The Melbourne Declaration on 100% Voluntary Non-remunerated Donation of Blood and Blood Components*. https://www.who.int/publications/m/item/who-global-consultation-100-voluntary-non-remunerated-blood-donation-of-blood-and-blood-components (WHO, 2009).

32. European Centre for Disease Prevention and Control (ECDC). *Facilitating COVID-19 Vaccination Acceptance and Uptake in the EU/EEA* (ECDC, 2021).

33. Haushofer, J. & Metcalf, C. J. E. Which interventions work best in a pandemic? *Science* **368**, 1063–1065 (2020).

34. Dai, H. et al. Behavioural nudges increase COVID-19 vaccinations. *Nature* **597**, 404–409 (2021).

35. Saccardo, S. et al. Assessing nudge scalability: two lessons from large-scale RCTs. *SSRN*, https://doi.org/10.2139/ssrn.3971192 (2022).

36. Bartoš, V., Bauer, M., Cahlíková, J. & Chytilová, J. Communicating doctors' consensus persistently increases COVID-19 vaccinations. *Nature* **606**, 542–549 (2022).

37. Oza, A. Studies probe how payouts affect U.S. vaccination rates. *Science* **373**, 611 (2021).

38. Loewenstein, G. & Cryder, C. Why paying people to be vaccinated could backfire. *The New York Times* https://www.nytimes.com/2020/12/14/upshot/covid-vaccine-payment.html (2020).

39. Persad, G. & Emanuel, E. J. Ethical considerations of offering benefits to COVID-19 vaccine recipients. *JAMA* **326**, 221–222 (2021).

40. Thaler, R. More than nudges are needed to end the pandemic. *The New York Times* https://www.nytimes.com/2021/08/05/business/vaccine-pandemic-nudge-passport.html (2021).

41. Volpp, K. G., Loewenstein, G. & Buttenheim, A. M. Behaviorally informed strategies for a national COVID-19 vaccine promotion program. *JAMA* **325**, 125–126 (2021).

42. Brewer, N. T., Chapman, G. B., Rothman, A. J., Leask, J. & Kempe, A. Increasing vaccination: putting psychological science into action. *Psychol. Sci. Public Interest* **18**, 149–207 (2018).

43. Gneezy, U., Meier, S. & Rey-Biel, P. When and why incentives (don't) work to modify behavior. *J. Econ. Perspect.* **25**, 191–210 (2011).

44. Lacetera, N., Macis, M. & Slonim, R. Rewarding volunteers: a field experiment. *Manage. Sci.* **60**, 1107–1129 (2014).

45. Goette, L. & Stutzer, A. Blood donations and incentives: evidence from a field experiment. *J. Econ. Behav. Organ.* **170**, 52–74 (2020).

46. Becker, G. S. & Elías, J. J. Introducing incentives in the market for live and cadaveric organ donations. *J. Econ. Perspect.* **21**, 3–24 (2007).

47. Berlin, I. et al. Financial incentives for smoking cessation in pregnancy: multicentre randomised controlled trial. *Br. Med. J.* **375**, e065217 (2021).

48. Milkman, K. L. et al. Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).

49. Barankay, I. et al. Effect of patient financial incentives on statin adherence and lipid control: a randomized clinical trial. *JAMA Netw. Open* **3**, e2019429 (2020).

50. Halpern, S. D. et al. Effectiveness and ethics of incentives for research participation: 2 randomized clinical trials. *JAMA Intern. Med.* **181**, 1479–1488 (2021).

51. Hyun, I. Fair payment or undue inducement? *Nature* **442**, 629–630 (2006).

52. Facciorusso, A., Demb, J., Mohan, B. P., Gupta, S. & Singh, S. Addition of financial incentives to mailed outreach for promoting colorectal cancer screening: a systematic review and meta-analysis. *JAMA Netw. Open* **4**, e2122581 (2021).

53. Berger, R. L. & Hsu, J. C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.* **11**, 289–319 (1996).

54. Lakens, D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Per. Sci.* **8**, 355–362 (2017).

55. Rabb, N. et al. Evidence from a statewide vaccination RCT shows the limits of nudges. *Nature* **604**, E1–E7 (2022).

56. Jacobson, M., Chang, T. Y., Shah, M., Pramanik, R. & Shah, S. Can financial incentives and other nudges increase COVID-19 vaccinations among the vaccine hesitant? A randomized trial. *Vaccine* **40**, 6235–6242 (2022).

57. Thirumurthy, H., Milkman, K. L., Volpp, K. G., Buttenheim, A. M. & Pope, D. G. Association between statewide financial incentive programs and COVID-19 vaccination rates. *PLoS ONE* **17**, e0263425 (2022).

58. Brewer, N. T. et al. Incentives for COVID-19 vaccination. *Lancet Reg. Health Americas* **8**, 100205 (2022).

59. Acharya, B. & Dhakal, C. Implementation of state vaccine incentive lottery programs and uptake of COVID-19 vaccinations in the United States. *JAMA Netw. Open* **4**, e2138238 (2021).

60. Barber, A. & West, J. Conditional cash lotteries increase COVID-19 vaccination rates. *J. Health Econ.* **81**, 102578 (2022).

61. Milkman, K. L. et al. A citywide experiment testing the impact of geographically targeted, high-pay-off vaccine lotteries. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-022-01437-0 (2022).

62. Robertson, C., Schaefer, K. A. & Scheitrum, D. Are vaccine lotteries worth the money? *Econ. Lett.* **209**, 110097 (2021).

63. Campos-Mercade, P., Meier, A. N., Schneider, F. H. & Wengström, E. Prosociality predicts health behaviors during the COVID-19 pandemic. *J. Public Econ.* **195**, 104367 (2021).

64. Andersson, O., Campos-Mercade, P., Meier, A. N. & Wengström, E. Anticipation of COVID-19 vaccines reduces willingness to socially distance. *J. Health Econ.* **80**, 102530 (2021).

65. Royer, H., Stehr, M. & Sydnor, J. Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a Fortune-500 company. *Am. Econ. J. Appl. Econ.* **7**, 51–84 (2015).

66. Ngo, S., Kim, A. S. & Chiong, W. Evidence for the ethics of incentivizing clinical trial enrollment? *JAMA Intern. Med.* **181**, 1488–1489 (2021).

67. Ambuehl, S., Niederle, M. & Roth, A. E. More money, more problems? Can high pay be coercive and repugnant? *Am. Econ. Rev.* **105**, 357–360 (2015).

68. Elías, J. J., Lacetera, N. & Macis, M. Paying for kidneys? A randomized survey and choice experiment. *Am. Econ. Rev.* **109**, 2855–

2888 (2019).

69. Rosenthal, R. *Experimenter Effects in Behavioral Research* (Appleton-Century-Crofts, 1969).

70. Haaland, I., Roth, C. & Wohlfart, J. Designing information provision experiments. *J. Econ. Lit.* (2021).

Acknowledgements

We are very grateful for the support of the Public Health Agency of Sweden regarding the linkage of the trial data with administrative vaccination records. We thank M. Alsan, S. Ambühl, B. Bartling, S. Bowles, A. Buttenheim, E. Fehr, U. Gneezy, J. Haushofer, D. Huffman, N. Lacetera, E. Largent, G. Loewenstein, M. Macis, M. Maréchal, R. Perez-Truglia, P. Rey-Biel, J. Savulescu, A. Stutzer, K. Volpp and R. Weber for feedback. We acknowledge funding from the Swiss National Science Foundation grant no. 100018_185176 (F.H.S.), CEBI coming from the Danish National Research Foundation grant no. DNRF134 (P.C.-M.), the Chazen Institute for Global Business at Columbia Business School (S.M.), the Columbia Business School (S.M.), the Booth School of Business, University of Chicago (D.P.), the Riksbankens Jubileumsfond (E.W.) and the Swiss National Science Foundation grant no. PZ00P1_201956 (A.N.M.).

Funding

Open access funding provided by Lund University.

Author information

Author notes

1. These authors contributed equally: Florian H. Schneider, Pol Campos-Mercade, Armando N. Meier

Authors and Affiliations

1. University of Zurich, Zürich, Switzerland

Florian H. Schneider

2. CESifo, Munich, Germany

Florian H. Schneider

3. Lund University, Lund, Sweden

Pol Campos-Mercade & Erik Wengström

4. University of Copenhagen, Copenhagen, Denmark

Pol Campos-Mercade

5. Columbia Business School, New York, NY, USA

Stephan Meier

6. University of Chicago Booth School of Business, Chicago, IL, USA

Devin Pope

7. National Bureau of Economic Research, Cambridge, MA, USA

Devin Pope

8. Hanken School of Economics, Helsinki, Finland

Erik Wengström

9. Unisanté and Lausanne Center for Health Economics, Behavior, and Policy (LCHE), University of Lausanne, Lausanne, Switzerland

Armando N. Meier

10. Faculty of Business and Economics, University of Basel, Basel, Switzerland

Armando N. Meier

Contributions

F.H.S., P.C.-M. and A.N.M. are equal lead authors who conceived and led the study. All authors contributed to conceptualization, methodology, experiments, data analysis, funding and writing.

Corresponding authors

Correspondence to Florian H. Schneider, Pol Campos-Mercade or Armando N. Meier.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Sebastian Calonico and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Raw treatment effects of offering financial incentives for first-dose uptake on further COVID-19

vaccination, other health behaviours, morals and civic responsibility, perceived safety, efficacy and trust, and other concerns.

The figure is based on RCT data linked to comprehensive survey data and population-wide Swedish administrative data capturing each vaccination in Sweden. The figure shows the raw treatment effects (no control variables included) of the financial incentives condition relative to the control condition. The blue dots indicate the estimated impact in standard deviations on the respective variables; all outcomes are defined exactly as pre-registered. Error bars represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey lines indicate the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions across datasets are as follows: administrative data, *n* incentives = 1,132, *n* control = 3,888; first survey data, *n* incentives = 726, *n* control = 2,512; second survey data, *n* incentives = 606, *n* control = 2,100.

Extended Data Fig. 2 Regression-estimated effects of informing US residents about state vaccination incentive programmes on awareness.

The figure is based on experimental data from a general population sample of US residents in 12 states that offered incentive programmes for COVID-19 vaccination. The figure shows regression-estimated effects of the incentives condition (informing participants about the existence of incentive programmes in their state) relative to the control condition. The blue dots indicate the estimated impact in standard deviations on the general awareness about state incentive programmes and the awareness of the specific state incentive programme the participants were informed about. Error bars represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey line indicates the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions are *n* incentives = 1,521 and *n* control = 1,541.

[Extended Data Fig. 3 Raw treatment effects of informing US residents about state vaccination incentive programmes on further COVID-19 vaccination, other health behaviours, morals and civic responsibility, and perceived safety, efficacy and trust.]

The figure is based on experimental data from a general population sample of US residents in 12 states that offered incentive programmes for COVID-19 vaccination. The figure shows the raw effects (no control variables included) of the incentives condition (informing participants about the existence of incentive programmes in their state) relative to the control condition. The blue dots indicate the estimated impact in standard deviations on the general awareness that the state had an incentive programme in place and on the awareness about the specific state incentive programmes participants were informed about. Error bars represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey lines indicate the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions are *n* incentives = 1,521 and *n* control = 1,541.

[Extended Data Fig. 4 Regression-estimated effects of informing US residents about state vaccination incentive programmes on single items of indices.]

The figure is based on experimental data from a general population sample of US residents in 12 states that offered incentive programmes for COVID-19 vaccination. The figure shows regression-estimated effects of the incentives condition (informing participants about the existence of incentive programmes in their state) relative to the control condition. The blue dots indicate the estimated impact in standard deviations on the respective variables. Error bars represent 95% confidence intervals (two-sided CI: mean ± 1.96 s.e.) from OLS regressions with heteroscedasticity-robust standard errors. The dashed grey lines indicate the threshold for small effect sizes of 0.2 standard deviations (Cohen's *d*). The sample sizes for the control and incentives conditions are *n* incentives = 1,521 and *n* control = 1,541.

Extended Data Table 1 Heterogeneous treatment effects based on vaccine hesitancy on second-dose uptake
Extended Data Table 2 Heterogeneous treatment effects based on vaccine hesitancy on morals and civic responsibility, perceived safety and efficacy, and trust
Extended Data Table 3 Heterogeneous treatment effects based on sociodemographics on second-dose uptake
Extended Data Table 4 Heterogeneous treatment effects based on sociodemographics on morals and civic responsibility, perceived safety and efficacy, and trust

Supplementary information

Supplementary Information

This file contains Supplementary Methods, Supplementary Tables and Figures, and Questionnaires.

Reporting Summary

Source data

Source Data Fig. 1

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 4

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give

appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Schneider, F.H., Campos-Mercade, P., Meier, S. *et al.* Financial incentives for vaccination do not have negative unintended consequences. *Nature* **613**, 526–533 (2023). https://doi.org/10.1038/s41586-022-05512-4

- Received: 01 March 2022
- Accepted: 02 November 2022
- Published: 11 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05512-4

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

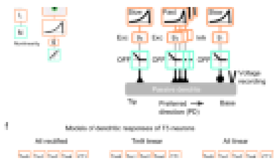
Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05512-4

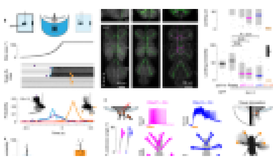
| Section menu | Main menu |



The physiological basis for contrast opponency in motion computation in Drosophila

17 August 2021

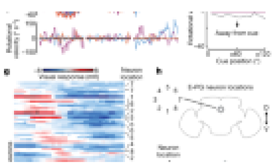
Giordano Ramos-Traslosheros & Marion Silies



State-dependent decoupling of sensory and motor circuits underlies behavioral flexibility in Drosophila

10 June 2019

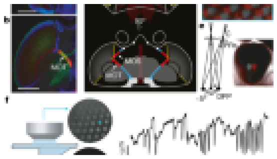
Jan M. Ache, Shigehiro Namiki, … Gwyneth M. Card



Sensorimotor experience remaps visual input to a heading-direction network

20 November 2019

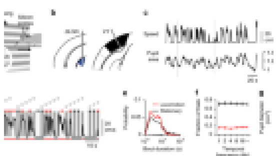
Yvette E. Fisher, Jenny Lu, … Rachel I. Wilson



[Muscles that move the retina augment compound eye vision in Drosophila]

26 October 2022

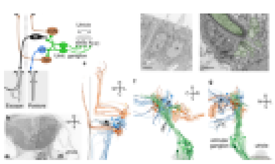
Lisa M. Fenk, Sofia C. Avritzer, … Gaby Maimon



[Locomotion modulates specific functional cell types in the mouse visual thalamus]

19 November 2018

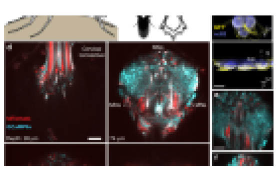
Çağatay Aydın, João Couto, … Vincent Bonin



[Organization of the gravity-sensing system in zebrafish]

27 August 2022

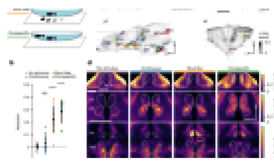
Zhikai Liu, David G. C. Hildebrand, … Martha W. Bagnall



Imaging neural activity in the ventral nerve cord of behaving adult Drosophila

22 October 2018

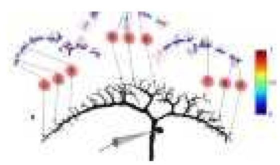
Chin-Lin Chen, Laura Hermans, … Pavan Ramdya



Visual recognition of social signals by a tectothalamic neural circuit

13 July 2022

Johannes M. Kappel, Dominique Förster, … Johannes Larsch



Non-uniform weighting of local motion inputs underlies dendritic computation in the fly visual system

10 April 2018

Ohad Dan, Elizabeth Hopp, … Idan Segev

- Article
- Open Access
- Published: 04 January 2023

Synaptic gradients transform object location to action

- Mark Dombrovski,
- Martin Y. Peek,
- Jin-Yong Park,
- Andrea Vaccari,
- Marissa Sumathipala,
- Carmen Morrow,
- Patrick Breads,
- Arthur Zhao,
- Yerbol Z. Kurmangaliyev,
- Piero Sanfilippo,
- Aadil Rehan,
- Jason Polsky,
- Shada Alghailani,
- Emily Tenshaw,
- Shigehiro Namiki,
- S. Lawrence Zipursky &
- …
- Gwyneth M. Card

Show authors

Nature **volume 613**, pages 534–542 (2023)

- 8812 Accesses
- 64 Altmetric
- Metrics details

Abstract

To survive, animals must convert sensory information into appropriate behaviours[1,2]. Vision is a common sense for locating ethologically relevant stimuli and guiding motor responses[3,4,5]. How circuitry converts object location in retinal coordinates to movement direction in body coordinates remains largely unknown. Here we show through behaviour, physiology, anatomy and connectomics in *Drosophila* that visuomotor transformation occurs by conversion of topographic maps formed by the dendrites of feature-detecting visual projection neurons (VPNs)[6,7] into synaptic weight gradients of VPN outputs onto central brain neurons. We demonstrate how this gradient motif transforms the anteroposterior location of a visual looming stimulus into the fly's directional escape. Specifically, we discover that two neurons postsynaptic to a looming-responsive VPN type promote opposite takeoff directions. Opposite synaptic weight gradients onto these neurons from looming VPNs in different visual field regions convert localized looming threats into correctly oriented escapes. For a second looming-responsive VPN type, we demonstrate graded responses along the dorsoventral axis. We show that this synaptic gradient motif generalizes across all 20 primary VPN cell types and most often arises without VPN axon topography. Synaptic gradients may thus be a general mechanism for conveying spatial features of sensory information into directed motor outputs.

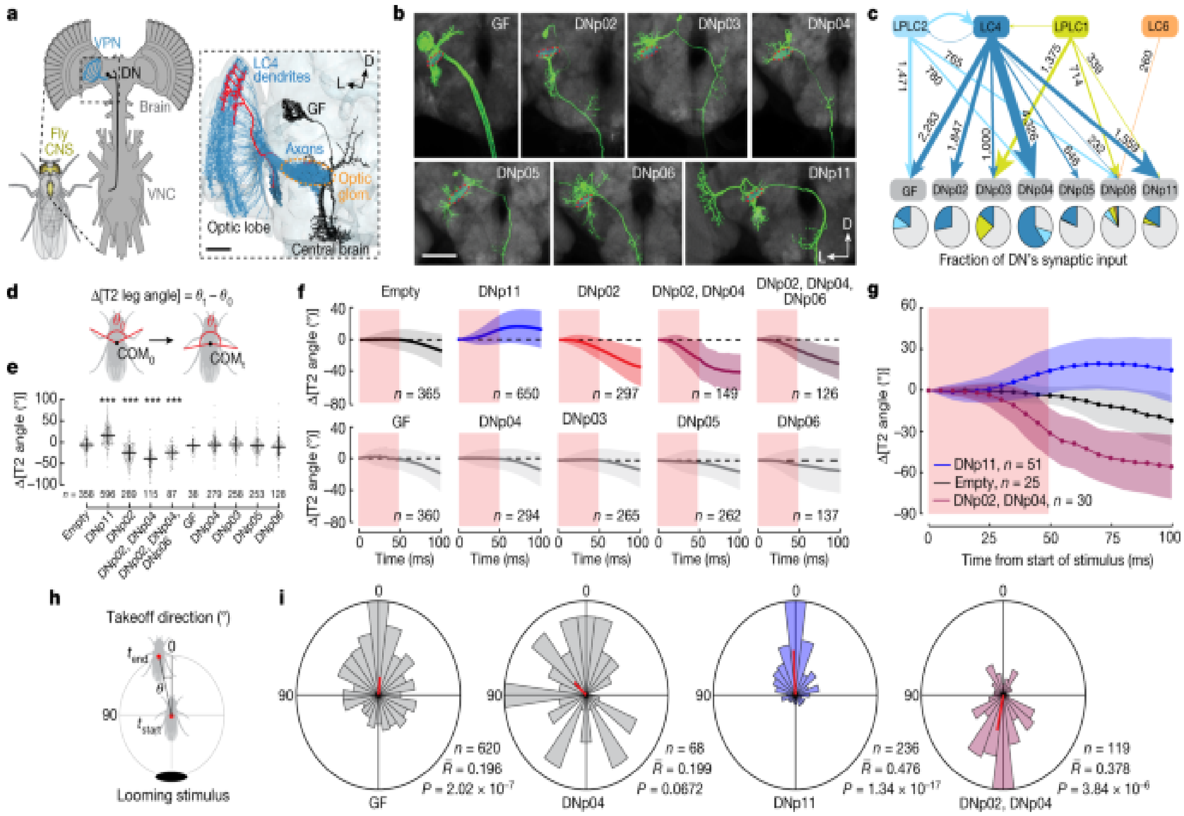
Main

To catch a ball, turn when called or pick up a cup, our brains must direct not just what to do, but where to do it. Inherent to this process is a 'sensorimotor transformation'[2,8,9] in which an object's location detected in sensory space, such as the position on the retina, is converted into movement direction in motor coordinates, such as the direction of limb or joint angle changes. There is considerable evidence that topographically organized brain regions in a wide range of species encode the location and identity of visual objects[10,11,12,13]; however, how neural connectivity patterns convey such information to downstream premotor networks, and how developmental programs specify this connectivity, remains poorly understood.

In *Drosophila*, VPNs that have dendrites in the optic lobe and axon terminals in the central brain detect ethologically relevant visual features,

such as small-object motion or looming of dark objects[6,7,14,15,16,17], and are close to the sensorimotor interface. Multiple VPN types initiate visually guided behaviours[6,18,19,20,21], and some VPN types synapse directly onto a subset of the ≈500 premotor descending neurons (DNs) per hemibrain whose activation drives distinct motor actions[22,23,24]. There are 20–30 different types of VPN, each a population of 20–200 neurons per hemibrain (Fig. 1a), with small receptive fields (20–40°) that together cover visual space[6,15,16]. VPN dendrites in the optic lobe thus form a topographic map of visual space, and object location on the fly's retina is theoretically encoded by which VPN neurons within a given type are excited. However, it has been unclear whether, and how, this spatial information is passed to downstream partners because the axons of all VPNs within a given type terminate in narrow, distinct glomeruli within the central brain (Fig. 1a) with little[25] or no[6,15,26,27] observable topography at the light-microscopy level. Yet several VPN cell types have been associated with direction-specific behaviours, including backing up and turning, escaping looming stimuli from different directions, collision avoidance and, in flight, saccade turns away from a visual stimulus[6,28,29,30]. Here we examine how direction-specific visual information is transformed onto downstream premotor networks by exploring the VPN-to-postsynaptic partner interface using electron microscopy (EM), light microscopy, physiology and behaviour.

Fig. 1: LC4 VPNs pass looming location information to DNs that mediate forward or backward escape takeoffs.



a, VPNs with retinotopically arranged dendrites in the lobula neuropil of the fly optic lobe have axon terminals in cell-type-specific optic glomeruli in the central brain. Dendrites of >50 postsynaptic neurons typically innervate each optic glomerulus. Inset: EM-based reconstructions (hemibrain connectome[27]) of 71 LC4 VPNs (blue), a single LC4 neuron (red) and LC4 postsynaptic partner, GF DN (black). VNC, ventral nerve cord; D, dorsal; L, lateral; glom., glomerulus. Scale bar, 20 µm. **b**, Confocal projections of GFP (green) expression in seven DNs innervating the LC4 glomerulus (red dashed line). Grey, brain neuropils. Images adapted from ref. [24], CC BY 4.0 ($n = 4$ brains for each DN). Scale bar, 50 µm. **c**, Synaptic connectivity from looming-sensitive VPN cell types onto seven DNs based on the hemibrain connectome. Arrow width is proportional to synapse number. Pie charts indicate proportion of a given DN's inputs from each looming-sensitive VPN cell type. **d**, Forward–backward postural shifts in response to DN photostimulation; quantified as Δ[T2 leg angle], the change in angle between the middle jumping legs and COM. **e**, Δ[T2 leg angle] 75 ms after the onset of 50-ms photostimulation. Points, individual flies; error bars, s.d.; one-way analysis of variance (ANOVA), Dunnett's test, ***$P < 0.001$, exact P values in Supplementary Table 1. **f**, Δ[T2 leg angle]

time courses from machine-learning-tracked data; red shaded area, photostimulation period. **g**, Δ[T2 leg angle] for a subset of manually annotated flies. In **f**,**g**: lines, mean; shading, s.d. **h**, Takeoff direction is COM movement direction between onset of middle leg extension and takeoff. **i**, Polar histograms of optogenetically activated takeoff direction. Red line, circular mean; *n*, number of flies tested; $\bar{R}$, mean vector length; *P*, Hodges–Ajne test for angular uniformity.

Source data

Neural control of looming escape direction

Looming visual cues indicate an impending collision or predator attack and drive rapid escape actions in most visual animals[31,32]. Flies orient their escape takeoff away from the direction of a looming stimulus[28,33]. Several *Drosophila* VPN types respond to looming stimuli[6,16,33,34], in particular LC4, a population of about 60 neurons per hemibrain, whose activation is critical for fast escape takeoffs through direct synapses onto the giant fibre (GF) DN[35] (Fig. 1a). To investigate the control of escape direction, we measured fly responses to three different directions of looming using the FlyPEZ[33] automated assay and machine-learning-based automated tracking (Extended Data Fig. 1a). Flies moved their centre of mass (COM) away from the stimulus direction (Extended Data Fig. 1a), and takeoffs were generally[33] away from the stimulus (Extended Data Fig. 1b). As previously suggested[28], we found takeoff direction arose from pre-takeoff postural shifts of a fly's COM relative to its middle pair of legs (Δ[T2 leg angle]; Extended Data Fig. 1c,d), which power the takeoff jump. This indicates that object location encoded by looming-sensitive VPNs, such as LC4, is passed downstream.

GF activation does not drive postural adjustments[36] and is not expected to control the escape takeoff direction. LC4 axons, however, overlap with dendrites of nine other DNs[24] (here called LC4-DNs). To examine whether LC4-DNs control takeoff direction, we focused on seven for which we had DN-specific genetic driver lines[24] (Fig. 1b). Analysis of the *Drosophila* 'hemibrain connectome', reconstructed from EM data[27], confirmed that

these DNs receive direct visual input from looming-sensitive VPNs, and (except for DNp06) a substantial portion of this is from LC4 (Fig. 1c) with four of them (DNp04, GF, DNp02 and DNp11) among the top 10 downstream partners of LC4 (ref. [27]). We optogenetically activated each DN, as well as two 'combination' lines targeting either two or three LC4-DNs together, and analysed the resulting behaviour with high-speed video[33]. GF activation produced takeoff rates of greater than 90% (refs. [33,34,35,36]). Only DNp04, DNp11 and combination line activation increased takeoff rates significantly compared to that of controls (Extended Data Fig. 1e,f and Supplementary Table 1), albeit with rates lower than that for GF activation (that is, 15–40% versus >90%), suggesting that natural threats may simultaneously activate multiple LC4-DNs to drive downstream escape motor circuits. DNp04- and DNp11-activated takeoffs were almost exclusively 'long-mode', in which the wings are raised before the takeoff jump, whereas GF activation produced 'short-mode' escapes without prior wing-raising as previously described[36] (Extended Data Fig. 1g,h and Supplementary Table 1). Combination line activation drove primarily long-mode takeoff, but did also unexpectedly produce many short-mode takeoffs, which are thought to rely on GF activation. Taken together with the findings of our previous work[37], this mixed result indicates either that the combination of DNp02, DNp04 and DNp06 inputs to the GFs, or that these DNs are not naturally co-activated with the strong intensity of optogenetic activation.

To evaluate whether any of these DNs triggered postural adjustments critical for escape directionality, we tracked 11 body points using Animal Part Tracker software (Branson Lab, see Methods) and created a metric for postural shift (Fig. 1d). DNp11 activation drove flies to lean forwards, whereas activation of DNp02 (including combinations of DNp02 and DNp04 or DNp02, DNp04 and DNp06) promoted backward leaning (Fig. 1e–g and Supplementary Videos 1 and 2). We next assessed whether these induced postural shifts led to directional takeoffs (Fig. 1h,i). Activation of DNp11 evoked forward takeoffs (Fig. 1i), whereas activation of DNp02 and DNp04 together evoked a strong bias towards backward takeoffs (Fig. 1i). As activation of DNp04 alone resulted in omnidirectional takeoffs (Fig. 1i), we reasoned that DNp02 was the main contributor to the movements leading to backward takeoff. The weak forward takeoff bias from GF activation

probably results from the average resting posture of the fly, which was previously observed to have the COM slightly in front of the T2 legs[28].

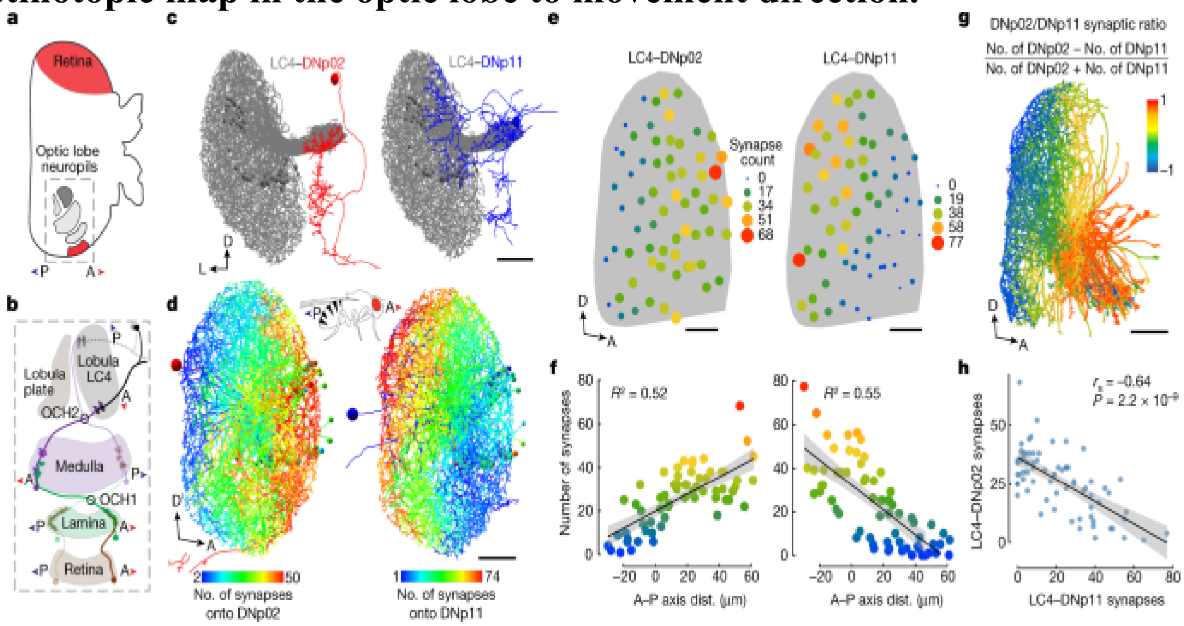
To further test whether DNp02 and DNp11 contribute to directional control during looming-evoked escape, we silenced each DN by selectively expressing Kir2.1, an inwardly rectifying potassium channel, and then measured responses to frontal (0°) or rear (180°) looming stimuli (Extended Data Fig. 2). DNp02-silenced flies took off normally (forwards) in response to rear stimuli but showed significant impairment in their ability to take off backwards in response to frontal stimuli—on average most DNp02-silenced flies took off forwards, directly towards the stimulus. This is consistent with the activation of DNp02 driving a backward postural shift, and supports a critical role for DNp02 in the postural adjustments that control backward takeoffs. Notably, flies in which DNp11 was silenced had a similar phenotype—these flies took off forwards in response to both frontal and rear looming stimuli. This could indicate that more DNs, possibly with interconnections, are involved in the control of forward takeoffs than backward ones, and also probably reflects the bias of the fly to jump forwards if no postural adjustment is made from the common resting posture. We conclude that, as flies with either DNp11 or DNp02 inactivated did not respond with normal takeoff directions to anterior or posterior looming stimuli, both DNs contribute to directional control of the fly's natural escape behaviour.

EM reveals LC4-to-DN synaptic gradients

We next sought to determine how LC4 neurons differentially convey the spatial location of the looming stimulus to DNp11 and DNp02 (Fig. 2a,b). In the right hemisphere of a complete serial section transmission EM dataset, we traced all LC4 neurons, DNp02 and DNp11 (FAFB dataset[38]; Fig. 2c) and marked synapses between LC4 neurons and each DN. We found a wide range (1 to 75) in the number of synapses individual LC4 neurons made with a given DN (Extended Data Fig. 3a). We next investigated whether LC4 neurons that synapsed more with DNp11 or DNp02 had dendrites located in a particular region of the lobula neuropil. We visualized the LC4 dendrites in the lobula and coloured each neuron by the number of synapses it made with a given DN. This revealed antiparallel synaptic number gradients along the

lobula anterior–posterior (A–P) axis for DNp02 and DNp11 (Fig. 2d). By contrast, A–P gradients were not seen in LC4 connectivity onto the GF and DNp04 (Extended Data Fig. 3b,c). The same A–P gradient patterns with LC4 synapses onto DNp11 and DNp02 were seen in an EM dataset from a second brain (hemibrain)[27] (Fig. 2e–g). This was supported by a strong negative correlation between the number of synapses a given LC4 makes with DNp11 and with DNp02 (Fig. 2h). The orientation of these gradients corresponds to the backward- and forward-jumping motor outputs of DNp02 and DNp11, respectively.

Fig. 2: Synaptic number gradients between LC4 and DNs transform a retinotopic map in the optic lobe to movement direction.



a, Fly visual system (dorsal view). The A–P axis of the visual space is mapped onto the anatomical lateral–medial axis of the lobula neuropil. The outlined area is shown in **b**. **b**, Anterior and posterior visual inputs to LC4 neurons through two optic chiasms (OCHs). Images in **a**,**b** adapted from ref. [50], CC BY 4.0. **c**, DNp02 (red) and DNp11 (blue) dendrites receive input from LC4 neurons (grey) in the glomerulus formed by LC4 axon terminals. Shown are neuron skeletons (red and blue). Scale bar, 50 µm. **d**, LC4 dendrites in the lobula (lateral view) colour-coded according to the number of synapses their axons make onto DNp02 or DNp11. LC4–DNp02 and LC4–DNp11 synaptic gradients are antiparallel along the A–P axis of the visual space. Scale bar, 20 µm. All neurons in **c**,**d** are manually reconstructed

from the EM FAFB dataset. **e**, Antiparallel A–P gradients are also seen in the hemibrain connectome. Dots, two-dimensional (2D) lobula projections of dendritic centroids for individual LC4 neurons in the lobula weighted in size and colour by the number of synapses made by their axons onto DNp02 and DNp11. Scale bars, 25 μm. **f**, Regression of LC4-DN synaptic weights as a function of LC4 dendrite centroid location; colour as in **e**. Linear fit line overlaid. Error bands, s.e.m. **g**, Hemibrain connectome reconstruction of LC4 dendrites coloured on the basis of a normalized (−1 to 1) number of synapses each LC4 neuron forms with DNp02 and DNp11. Some anterior lobula dendrites exceed the EM volume and are not fully reconstructed. **h**, Correlation between the number of synapses each LC4 neuron ($n = 71$) makes with DNp02 and DNp11. r_s, Spearman's rank correlation coefficient. A, anterior; P, posterior; D, dorsal; L, lateral. Error band, s.e.m.

Source data

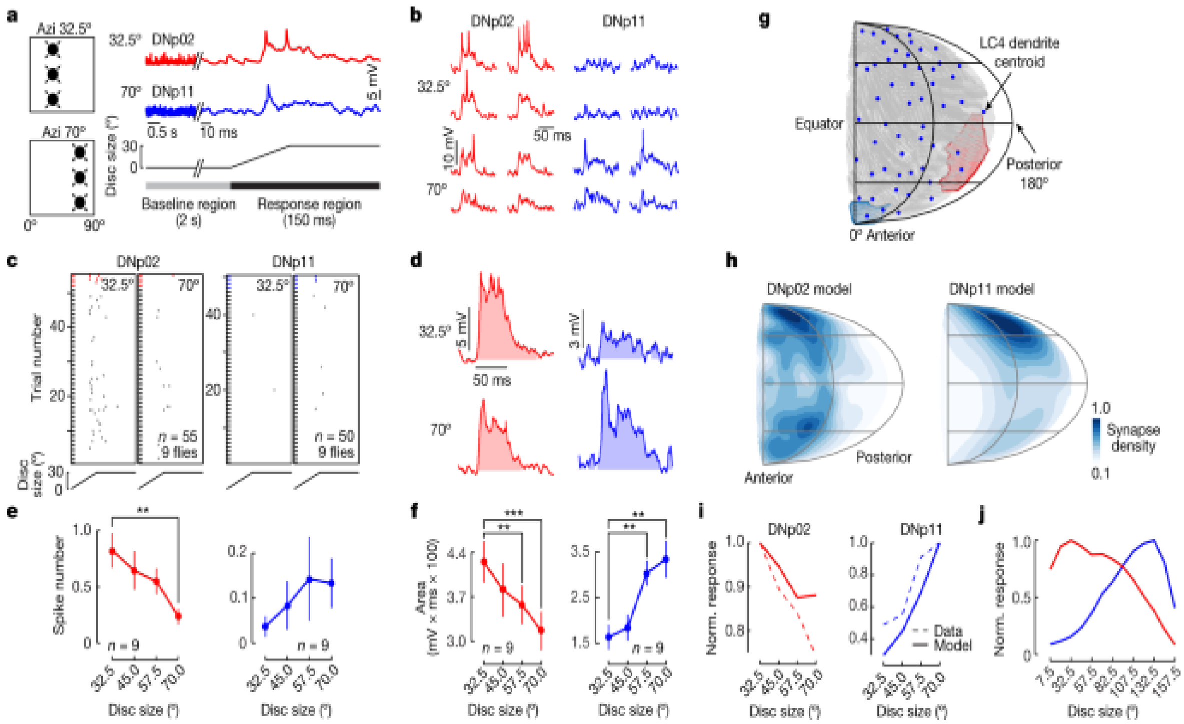
Taken together, the behaviour and connectomic data support a simple model: antiparallel synaptic gradients transform locally detected object location into oppositely directed behaviours. A frontward looming stimulus activates anterior LC4 neurons that provide relatively more drive to DNp02, which produces backward body movements generating a backward escape trajectory following co-activation with DNp04 or other escape pathways. For a stimulus looming from behind, posterior LC4 neurons become more active and drive DNp11 to generate forward postural shifts and a forward-directed takeoff.

Synaptic gradients are functional

The synapse gradient model is based on the assumption that synapse number correlates with connection strength. To directly test this, we carried out in vivo whole-cell patch-clamp recordings from DNp02, DNp11, DNp04 and the GF during visual looming stimulation at varying locations along the A–P axis of the visual space. We presented vertical arrays of small dark expanding discs at four different azimuthal locations ipsilateral to the targeted DN (Fig. 3a and Extended Data Fig. 4a). DNp02, DNp11, the GF and DNp04 all depolarized in response to looming, and all except the GF

produced action potentials (Fig. 3a–f and Extended Data Fig. 4b–j; see Methods for identification of action potentials).

Fig. 3: LC4 synaptic number gradients onto DNp02 and DNp11 are functional.



a, Whole-cell electrophysiological recordings of DNp02 (red) and DNp11 (blue) to looming stimuli at 32.5° (for DNp02) and 70° (for DNp11) in azimuth. Looming stimulus is an array of three discs expanding 0° to 30° diameter at 500° s^{-1}. Shown are representative traces from a single fly and stimulus. **b**, Representative responses from a single fly for 32.5° (top) and 70° (bottom) azimuth looming stimuli. **c**, Spike raster plots of DN responses during the 150 ms after looming onset. Coloured trials show the traces in **b**. **d**, Averaged response of the traces in **b** shows subthreshold depolarizing responses to looming stimuli. Shaded area, estimated depolarization from the baseline. **e**, Mean per-trial spike count across individual flies (from **c**). *n*, individual trials; $^{**}P < 0.01$. **f**, Pooled mean of integrated potentials across individual flies. *n*, individual trials. Repeated-measures one-way ANOVA, Dunnett's test. Error bars, s.e.m.; $^{**}P < 0.01$, $^{***}P < 0.001$, see Supplementary Table 1 for exact *P* values. **g**, Mollweide projection of estimated dendritic receptive fields for all 55 LC4 neurons in the FAFB EM dataset. Polygons are estimated visual fields of individual LC4 neurons

(example individual fields in red and blue). **h**, DNp02 and DNp11 LC4-receptive fields estimated on the basis of summed input from individual LC4 fields in **g**. **i**, In vivo whole-cell (dashed) and model-estimated (solid) DN responses to three-loom-array stimuli (solid). **j**, Estimated DNp02 and DNp11 responses to modelled three-loom-array stimuli across the whole visual hemifield, based on receptive fields in **h**.

Source data

DNp02 produced more action potentials in response to anterior, compared to more posterior, stimuli (44 versus 13 spikes across all trials), whereas DNp11 exhibited the opposite trend (Fig. 3b,c,e and Extended Data Fig. 4i). These trends were consistent for both individual (Fig. 3c and Extended Data Fig. 4c,i) and averaged (Fig. 3e) responses. By contrast, DNp04 produced bursts of action potentials without significant azimuth tuning (Extended Data Fig. 4c–f,i). In agreement with the action potential tuning curves, depolarizing membrane potentials in DNp02 were larger for more anterior azimuthal locations of the looming stimulus, whereas those for DNp11 were larger for more posterior looming locations (Fig. 3d,f and Extended Data Fig. 4j). For the GF, we did not see distinct tuning properties for the anterior–posterior location of the stimuli. DNp04 did show a trend towards stronger responses to anterior stimuli, although the responses were more variable than for DNp11 or DNp02 (Extended Data Fig. 4g,h,j).

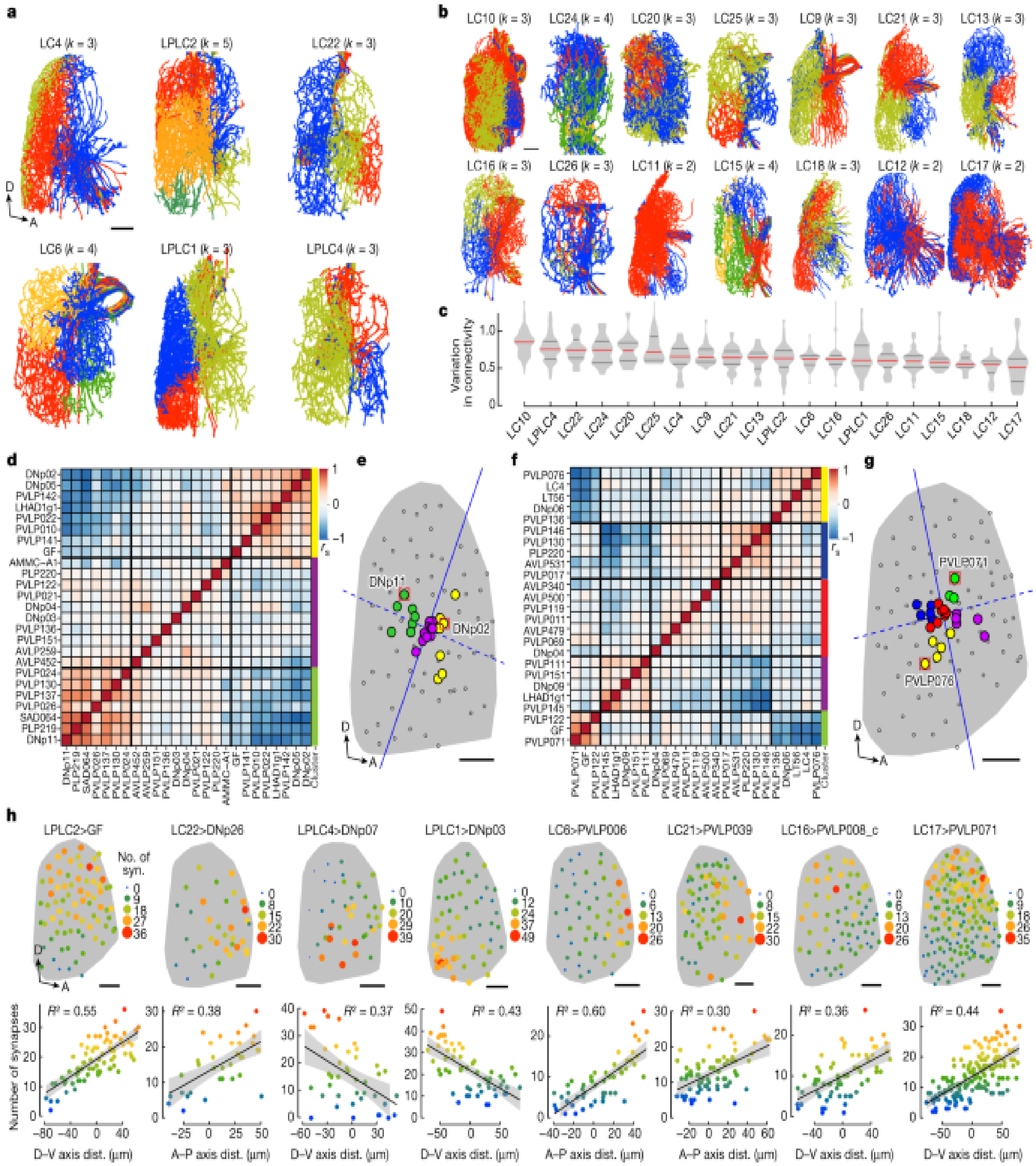
If synapse number correlates directly with input current drive to the postsynaptic cell, we should be able to predict the DN responses to looming stimuli at different azimuthal locations. To assess this, we used the EM data to make a model incorporating both the spatial profile of LC4 dendrites and the synaptic connectivity of LC4 axons with DNs. Main dendritic branches of all 55 LC4 neurons in the FAFB dataset[38] were mapped from lobula to eye coordinates following a previously established method[25] (Fig. 3g and Extended Data Fig. 5a–c). The normalized estimated responses to looming recapitulated the azimuthal tunings predicted by the synaptic gradients and matched the responses for all four DNs we measured (Fig. 3h,i and Extended Data Fig. 5d–f). We conclude that the synaptic numbers observed from EM data can be interpreted as functional synaptic weights.

We used this model to simulate responses to looming from azimuthal locations across the whole visual hemisphere, including those not possible in our physiology experiments. Our simulation showed strong antiparallel looming response profiles for DNp02 and DNp11 across nearly the whole visual hemifield (30°–130°), supporting the observed synaptic gradients as predictive of functional response profiles (Fig. 3j). Taken together, these results corroborate the model that anterior LC4 neurons provide stronger inputs to DNp02 in response to anterior stimuli whereas posterior LC4 neurons provide more drive to DNp11 in response to posterior stimuli in a graded fashion. This differential connectivity drives the backward (DNp02) or forward (DNp11) escape takeoffs away from looming threats.

Synaptic gradients are a common wiring motif

To address the question of whether visuomotor transformation through gradients of synapses is limited to just LC4 and DNp02 and DNp11 or whether it represents a general circuit wiring logic, we analysed the output connectivity patterns of 20 VPN cell types[6] using data from the hemibrain connectome[27]. First, we used principal component analysis and *k*-means analyses to cluster individual neurons within a VPN cell type on the basis of the similarity of their outputs (that is, the number of synapses they form onto the set of synaptic partners within their respective optic glomerulus; Extended Data Fig. 6a and Methods). Next, we colour-coded each cluster to visualize the relationship between a neuron's cluster identity and the spatial location of its dendrites in the lobula. A striking spatial separation of the clusters was found in most VPN cell types (Fig. 4a,b), revealing widespread differential synaptic connectivity, such that individual neurons within one VPN cell type elaborated quantitatively and qualitatively different outputs in the glomerulus depending on the location of their dendrites in the lobula (Fig. 4c and Extended Data Fig. 6b–g).

Fig. 4: Synaptic gradients are a general property of VPN output organization.



a,**b**, Connectivity-based *k*-means clustering of individual neurons within 20 VPN cell types (see Methods). Lateral views of VPN dendrites in the lobula (hemibrain connectome reconstructions). Individual cells within one VPN cell type are coloured by their cluster identity. Colours do not correspond between VPN types. Most VPNs exhibit distinct spatial separation (examples in **a**), but in some cases (LC12 and LC17 in **b**) there is no clear separation. Scale bars, 20 µm. **c**, Differential connectivity (number of synapses) across individual neurons within one VPN cell type. Measured for

20 VPN cell types and their postsynaptic partners that make at least 50 synapses total. Coefficients of variation in synapse number are averaged across all postsynaptic partners per VPN cell type. **d**, Matrix of pairwise correlations in synaptic connectivity between LC4 and its top 25 postsynaptic partners; ordered by hierarchical clustering as indicated by coloured side bars; r_s, Spearman's rank correlation coefficient. **e**, Topographic map of input centroids, weighted by number of synapses, for top 25 postsynaptic partners of LC4. Dark grey shading, lobula 2D projection; small open circles, centroids of 71 individual LC4 dendrites; coloured circles, weighted input centroids; solid blue line, median separation line; dashed blue line, projection line (see Methods). Red squares indicate centroids of DNp02 and DNp11. Scale bar, 25 µm. **f**,**g**, Similar analysis as in **d**,**e**, but for LPLC2. Red squares, centroids of neurons PVLP071 and PVLP076. Scale bar, 25 µm. **h**, Representative examples of synaptic gradients reflecting A–P and D–V axes of dendritic maps in multiple VPN cell types. syn., synapses. Scale bars, 25 µm (images 1 and 6–8) and 30 µm (images 2–5). A, anterior; P, posterior; D, dorsal; V, ventral. Error bands, s.e.m. See legend for Fig. 2c,d.

Source data

To investigate these properties in more detail, we analysed synaptic connectivity between two VPN cell types (LC4 and LPLC2) and the top 25 postsynaptic partners of each of them (Fig. 4d–g). Both VPN cells types are looming detectors and share some postsynaptic partners, including the GF[6,34,39]. For each VPN cell type, we first assessed the similarity of its outputs onto different postsynaptic neurons by measuring the pairwise correlation for all 300 possible pairs of its top 25 postsynaptic partners (similar to LC4 and DNp02 or DNp11 in Fig. 2h). The resulting matrices revealed that postsynaptic targets of LC4 and LPLC2 formed three and five connectivity-based clusters, respectively (Fig. 4d,f). Thus, different postsynaptic partners receive different patterns of input from the same VPN cell type. Next, to visualize the relationship between this differential input and VPN dendritic maps, we calculated weighted dendritic input centroids for each of the top 25 postsynaptic partners of LC4 and LPLC2, and measured pairwise distances between them (Extended Data Fig. 9a–h and Methods). These indicate spatial regions of the lobula providing the

most input to a given postsynaptic partner. The resulting topographic maps (Fig. 4e,g) revealed that all three connectivity-based clusters for LC4 clearly segregated along the A–P axis of the lobula (Fig. 4e). By contrast, two out of five clusters for LPLC2 segregated along the A–P axis of the lobula, two segregated along the D–V axis, and one cluster had no spatial bias (that is, neurons from this cluster receive uniform input from all LPLC2 neurons; Fig. 4g). Notably, both the numbers and topographic positions of these clusters largely match the results of *k*-means analysis for both VPN cell types (Fig. 4a).

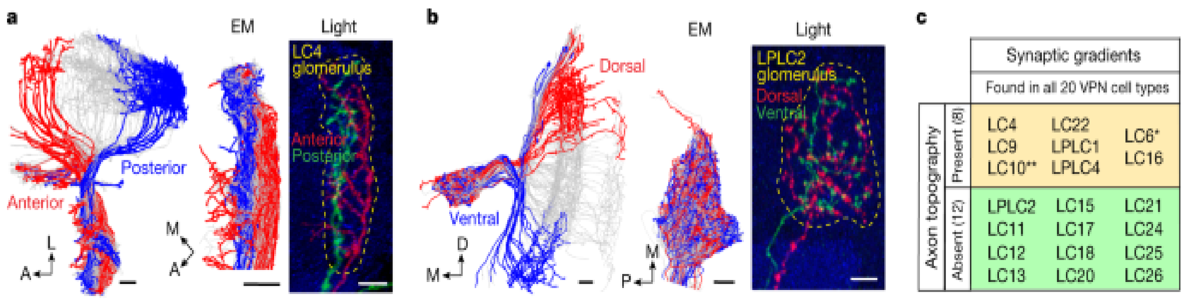
These examples illustrate how the topographic map of VPN dendritic inputs in the optic lobe is converted into maps of graded synaptic weights in the optic glomerulus. We observed synaptic gradients reflecting both the A–P and D–V axes of the dendritic map across all 20 VPN cell types (Fig. 4h and Extended Data Fig. 7), analogous to those we originally found in the fly directional escape circuit (Fig. 2). The ethological relevance of some of these gradients may be deduced from the known function of postsynaptic neurons in the literature. For example, the D–V gradient from LPLC4 onto DNp07 may control landing behaviour[22] (Fig. 4h) and the A–P gradient from LPLC1 onto PLP219 (Extended Data Fig. 7) could regulate collision avoidance[29]. Thus, we propose that conversion of visual space coordinates into graded synaptic connectivity is a shared feature of VPN wiring.

Synaptic gradients with or without axon topography

Topographic arrangement of VPN axons would provide a simple mechanism for the development of synaptic gradients. Previous studies concluded that this was unlikely[6,15,25,26] (with an exception of LC10 (refs. [6,13]) and traces of topography in the LC6 (ref. [25]) glomerulus). Here we revisited this issue using EM data[27] and looked for axon topography corresponding to dendritic arrangement along either the A–P or D–V axis of the lobula. We found five additional VPN cell types (LC4, LC9, LC22, LPLC1 and LPLC4) that have axon terminals retaining rough A–P topography, and one (LC16) whose axons maintain traces of D–V topography (Fig. 5a, Extended Data Fig. 8a–e and Supplementary Videos 3 and 4). These observations were confirmed

using light microscopy and MultiColor FlpOut[40]: the axon terminals of sparsely labelled VPNs with dendrites in either the anterior or posterior lobula targeted distinct domains in their corresponding glomeruli and also exhibited differential morphology as assessed by EM and light microscopy (Extended Data Fig. 8g–j). No axon topography, however, was observed for most (12/20) VPN cell types (Fig. 5b and Extended Data Fig. 8f) at the resolution of our analysis. Therefore, synaptic gradients in these cases (Fig. 4h and Extended Data Fig. 7) must emerge by an alternative mechanism.

Fig. 5: Axon topography is present in some, but not most, optic glomeruli.



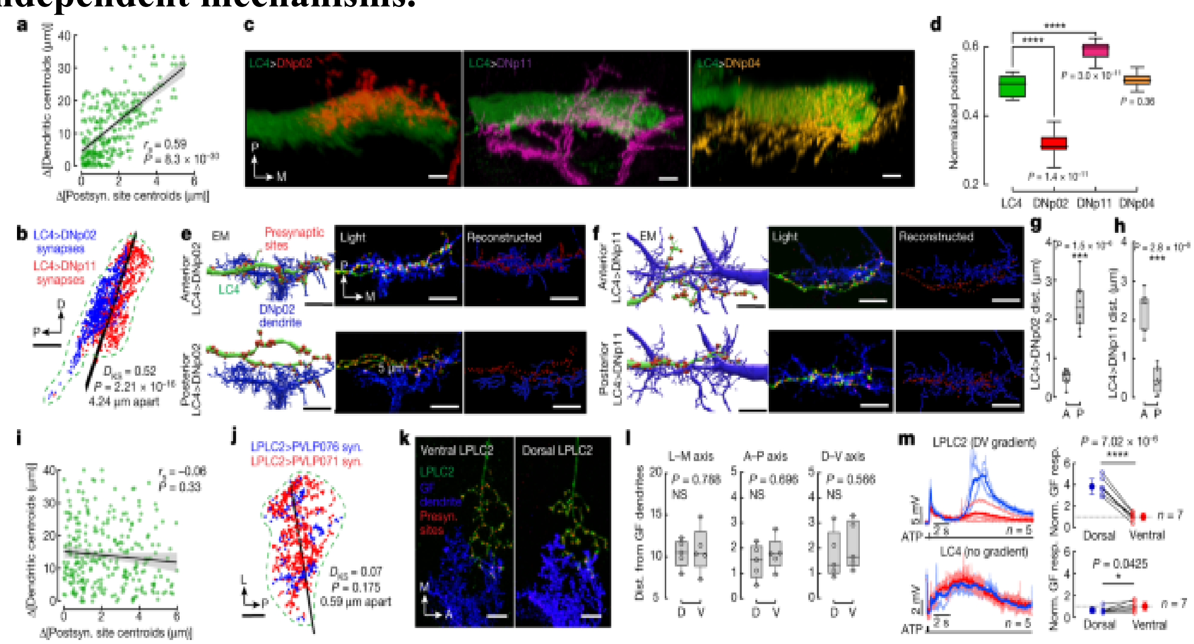
a, LC4 is a VPN cell type that retains axonal topography in optic glomerulus. Left, hemibrain connectome reconstructions of 15 anterior (red), 15 posterior (blue) and central LC4 cells (grey). Middle, EM reconstruction of axons in the LC4 glomerulus shows separation of anterior and posterior terminals. M, medial. Right, image of the LC4 glomerulus region with axon terminals of one anterior (red) and one posterior (green) cell labelled using MultiColor FlpOut and assessed using light microscopy ($n = 9$, all A–P pairs of individual clones from different brains exhibited reproducible axon terminal topography). Axonal projections form a topographic map in the glomerulus, corresponding to the location of their dendrites along the A–P axis of the lobula. Scale bars, 5 µm. **b**, LPLC2 is a VPN cell type without axonal topography. LPLC2 axon terminals do not form a topographic map along the D–V axis of the lobula as visualized from EM reconstruction (left and middle) and light microscopy (right, $n = 6$ pairs of clones). Scale bars, 5 µm. **c**, Relationship between synaptic gradients and topography of axon terminals for different VPN types (see Extended Data Fig. 8 for more examples). *LC6 retains coarse axonal retinotopy[25]. **LC10 was previously shown to have A–P axonal retinotopy[6,13]. A, anterior; P, posterior; D, dorsal; M, medial; L, lateral.

In summary, VPNs fall into two classes (Fig. 5c). In one, synaptic gradients correlate with axon topography within the glomerulus and in the other they do not.

DN dendrite location matches LC4 synaptic gradients

We focused on LC4 to understand how axon topography leads to the formation of synaptic gradients. We found that for the top 25 postsynaptic partners of LC4, the spatial distribution of postsynaptic sites in the LC4 glomerulus strongly correlated with the positions of LC4 dendrites in the lobula (Fig. 6a, Extended Data Fig. 9g,i and Methods). This is exemplified by DNp02 and DNp11 receiving anticorrelated inputs from LC4 axons (Figs. 2h and 4d,e) and having spatially segregated postsynaptic sites in the LC4 glomerulus (Fig. 6b and Extended Data Fig. 9i). Topographic mapping of the LC4 axon terminals alone cannot account for these patterns.

Fig. 6: Synaptic gradients in VPNs emerge through spatial and spatially independent mechanisms.



a–h, LC4: spatial mechanism. **a**, Relationship between dendritic map and spatial arrangement of synapses in the glomerulus for LC4 (for 300 pairs of top 25 postsynaptic (postsyn.) partners); r_s, Spearman's rank correlation

coefficient. Error bands, s.e.m. **b**, Location of DNp02 and DNp11 postsynaptic sites in the retinotopic LC4 glomerulus (green dashed outline); black line, separation plane; D_{KS}, two-tailed Kolmogorov–Smirnov test. Scale bar, 5 µm. **c**, Confocal projections of LC4 glomeruli and DN dendrites. Scale bars, 5 µm. **d**, Normalized DN dendritic centroid position within the LC4 glomerulus ($n = 12$ brains each; one-way ANOVA, Dunnett's test, versus LC4 glomerulus centroid, $****P < 0.0001$). **e**,**f**, Single anterior and posterior LC4 neurons with labelled presynaptic sites colocalized with DNp02 and DNp11 dendrites, alongside their EM reconstructions (bodyID 1907587934 and 1249932198). Scale bars, 5 µm. **g**,**h**, Distance (dist.) between dendrites of DNp02 (**g**) and DNp11 (**h**) and presynaptic sites in anterior versus posterior LC4 ($n = 8$ and 10 brains, respectively; two-tailed unpaired Welch's t-test, $***P < 0.001$). **i–m**, LPLC2: non-spatial mechanism. **i**, Relationship of input and output centroids (as in **a**) for LPLC2. **j**, Location of PVLP071 and PVLP076 postsynaptic sites in the LPLC2 glomerulus (statistics as in **a**,**b**), which lacks retinotopy. Scale bar, 5 µm. **k**, Single ventral and dorsal LPLC2 neurons with labelled presynaptic (presyn.) sites colocalized with the GF dendrite. Scale bars, 5 µm. **l**, Distances between presynaptic sites of single dorsal versus ventral LPLC2 neurons and GF dendrites, measured along three cardinal axes ($n = 5$ brains each; two-tailed unpaired Welch's t-test, NS, $P > 0.05$). **m**, GF depolarization responses from localized activation of dorsal versus ventral LPLC2 and LC4 neurons expressing the $P2X_2$ receptor. Left: representative GF responses ($n = 5$, one fly); individual (lighter-coloured lines) and averaged (darker lines) responses. Right: comparison of normalized average GF responses (resp.) to dorsal versus ventral VPN activation (two-tailed paired t-test; error bars, s.e.m., $*P \leq 0.05$, $****P < 0.0001$). Responses were averaged during the late response peak; see Extended Data Fig. 12c for quantification of the early peak. n, individual flies tested. A, anterior; P, posterior; D, dorsal; V, ventral; L, lateral; M, medial. All box plots show median and interquartile range.

Source data

To assess whether the spatial distribution of DN dendrites also contributes to differential connectivity, we mapped the positions of dendrites of different DN neurons within the LC4 glomerulus using light microscopy. DNp02 and

DNp11 dendrites occupy unique glomerular sub-compartments where axons of LC4 corresponding to anterior and posterior visual fields selectively terminate. By contrast, dendrites of DNp04, a postsynaptic neuron with no A–P synaptic gradient with LC4, arborize uniformly within the LC4 glomerulus (Fig. 6c,d and Extended Data Fig. 10a–c).

To map synapses at the light level, we used a modification of the STaR[41] method to visualize presynaptic sites in sparsely labelled LC4 neurons (Extended Data Fig. 10d,e) and assessed their proximity to DNp02 and DNp11 dendrites (Fig. 6e–h). The presynaptic sites of LC4 from the anterior lobula were much closer on average to the DNp02 dendrites than those from the posterior (Fig. 6e,g and Supplementary Videos 5 and 6). Conversely, DNp11 dendrites were closer to the presynaptic sites of LC4 from the posterior lobula.

In summary, LC4 utilizes a spatial wiring strategy to attain graded synaptic connectivity. A combination of topographic arrangement of LC4 axons and placement of DNp02 and DNp11 dendrites within different spatial domains in the glomerulus determines the directional specificity of the escape response to looming stimuli from different regions of the visual field.

Spatially independent synaptic gradients in LPLC2

The synaptic gradients elaborated by LPLC2 form in a fundamentally different way from those elaborated by LC4. Analysis of the top 25 postsynaptic partners of LPLC2 found no significant relationship between positions of LPLC2 dendrites in the lobula (that is, synaptic output specificity) and the spatial arrangement of synapses in the LPLC2 glomerulus (Fig. 6i and Extended Data Fig. 9h,j). For example, the postsynaptic neurons PVLP071 and PVPL076 have anticorrelated inputs from LPLC2 (Fig. 4f,g), yet their postsynaptic sites are intermingled in the LPLC2 glomerulus (Fig. 6j and Extended Data Fig. 9j).

We confirmed this principle by labelling presynaptic sites in axons of individual LPLC2 neurons with dendrites within the dorsal and ventral lobula and measuring the proximity of these presynaptic sites to the GF

dendrites (Fig. 6k and Extended Data Fig. 10f). No significant difference in distances was found (Fig. 6l) despite a marked difference in synapse counts (Fig. 4h). Thus, the spatial distribution of synapses in the LPLC2 glomerulus seems random. To assess this principle in a more systematic manner, we further analysed EM data (hemibrain) and measured the correlation between axo-dendritic overlap and synaptic counts for four topographic and four non-topographic VPNs and their postsynaptic partners (Extended Data Fig. 11). Our results strengthened the notion that VPNs utilize two qualitatively different wiring strategies to form synaptic gradients.

We next sought to assess whether the synaptic gradients of LPLC2 onto the GF were functionally significant (Fig. 4h). The dendrites of LPLC2 neurons expressing the $P2X_2$ receptor were locally activated by injection of ATP in the dorsal and ventral regions of the lobula, and the response in the GF was assessed using electrophysiological recordings (Extended Data Fig. 12a). GF responses following activation of dorsal LPLC2 were significantly stronger than those following ventral ATP injections. By contrast, little difference was seen in response following stimulation of dorsal versus ventral LC4 (also connected to the GF, but without a notable D–V synaptic gradient; Fig. 6m and Extended Data Fig. 12b,c).

In summary, functionally relevant graded synaptic connectivity of LPLC2 is established through a spatially independent mechanism.

Discussion

We took advantage of cell-type-specific genetic tools, behavioural and physiological analyses, and densely reconstructed neuron connectivity maps to examine a central brain sensory-to-motor interface at synaptic resolution. We showed that the transformation of object location from retinal to body coordinates is solved by gradients of synapses between spatially ordered visual-feature-detecting neurons (that is, VPNs) and movement-direction-specific premotor neurons (that is, DNs). We demonstrated that such numeric gradients produce functional synaptic weights and lead to predictable response differences in postsynaptic neurons that drive fly escape takeoffs correctly oriented away from looming threats. Individual cells within one VPN cell type are thus functionally heterogeneous with

connectivity profiles often as dissimilar as ones found between different neuron types. It is this continuous heterogeneity that converts visual stimuli into ethologically relevant behavioural responses.

We discovered behavioural roles for individual DNs (DNp02 and DNp11), and it may be tempting to consider these as command neurons for particular body movement directions. However, several observations suggest that they act instead as members of a larger DN group whose combined activity represents both the strength of the drive to takeoff and movement direction. First, when optogenetically activated alone no LC4-DN drove a high takeoff rate (25% takeoff rate maximum, all long-mode takeoffs). By contrast, activation of the command-like GF drove nearly 100% takeoff (all short-mode takeoffs). Second, activation of DN combinations (for example, DNp02 and DNp04 or DNp02, DNp04 and DNp06), increased takeoff rates significantly, although only up to about 40% takeoff. This suggests that co-activation of multiple DNs drives the long-mode takeoff and more DNs than we identified probably participate. Finally, whereas co-activation of DNp02 and DNp04 increased the backward shift of flies compared to activation of DNp02 alone, this shift was reduced by additional co-activation of DNp06. Thus, different DNs may 'vote' for movement in a particular direction and the resulting behaviour is the sum of these votes, much like the population activity in directionally selective motor cortex neurons correlates with movement direction in primates[42]. This mechanism could extend beyond forward and backward control if the left and right DNs of the same type, which would be differentially activated in the event of a looming stimulus from the side, also independently 'voted' for leftward or rightward body shifts, much like unilateral activity in DNg02 neurons correlates with left or right flight saccades in flying flies[43]. By this mechanism it would be plausible for the fly to obtain the ability to takeoff in any direction relative to its body, as has been observed in behavioural data.

Expanding our analysis to 20 different VPN cell types and their postsynaptic partners revealed synaptic gradients as a general property of visual feature detector output in *Drosophila*. Evidence consistent with a gradient motif has been observed at the sensorimotor interface of the cockroach cercal system, where input from directionally selective abdominal wind-sensitive hairs has graded effects on the response of downstream giant interneurons, which

drive escape[44]. Thus, synaptic number gradients may be a general principle for transmission of spatial information between sensory and motor networks.

VPNs guide innate visual behaviours of the fly, including looming-evoked backing or takeoff and small-object tracking[6,16]. We expect that the synaptic gradients we described here are specified by genetically hard-wired developmental processes, rather than through experience. In support of a developmental origin, we observed substantially the same LC4-DN gradients in EM volumes of two different fly brains[27,38]. The same wiring motif, however, could be present in more flexible areas of sensorimotor interface such as ellipsoid body ‘compass’ neurons[45] and would provide a simple mechanism for how learning-induced changes in numbers of synapses between neurons could result in different stimulus–behaviour pairings.

We identified two different circuit wiring strategies producing synaptic gradients in different VPN cell types. In the ‘spatial’ strategy, topographic mapping of VPN axon terminals organizes the optic glomerulus and is ‘read out’ by stereotypically positioned dendrites of different target neurons. Axonal topography may arise through age-dependent mechanisms as described for more peripheral regions of the fly visual system[46], or through graded expression of cell surface molecules (for example, Eph receptors and ephrins) as described in the vertebrate visual system[47]. Developmental mechanisms must act in parallel to target dendritic processes of different postsynaptic neuron types to discrete domains within the glomerulus.

Most VPN cell types we examined (12/20), however, did not show clear topographic organization of their axonal projections. Thus, in most cases, gradients emerge in the absence of spatial cues. Molecular heterogeneity within one cell type previously found in the fly visual system[48] and mouse visual cortex[49] may underlie such differential synaptic specificity. Future work should examine whether spatial gradients of molecular regulators instruct differential expression of cell adhesion and recognition molecules in VPNs, thereby transforming a retinotopic arrangement of dendritic arbours in the optic lobe into a graded distribution of synapses in the central brain.

Methods

Experimental model details

Flies were reared under standard conditions at 25 °C and 50% humidity with a 16-h light/8-h dark cycle on a standard cornmeal fly food. Male and female flies 3–5 days after eclosion were used for all experiments except if specified otherwise. Flies used for optogenetic activation experiments were raised on 0.2 mM retinal (Sigma R2500) food, and maintained on 0.4 mM retinal food as adults. These flies were kept in the dark in foil-covered vials until they were prepared for experiments. Supplementary Table 2 provides detailed descriptions of fly genotypes used in each experiment and origins of transgenic stocks.

Behavioural experiments

High-throughput takeoff assay

We tested escape responses of unrestrained flies using our previously developed FlyPEZ[33] system to automate fly behaviour experiments and collect large sample sizes necessary to quantitatively characterize differences in escape behaviour. In FlyPEZ, individual flies were released one at a time onto a 5 mm by 5 mm glass platform through an automated gate without undue perturbation, where they were targeted for visual or optogenetic stimulation. The fly position on the platform was tracked using a real-time tracking algorithm, which coordinated the triggering of a high-speed video camera and either looming stimulus or light stimulus. For visual stimulation, we used digital micromirror device projectors running at a refresh rate of 360 Hz, controlled by MATLAB using the Psychophysics Toolbox. Dark looming discs expanding from 10° to 180° at an elevation of 45° and azimuth of 0°, 90° or 180° ± 22.5° relative to the fly head position were presented on a 7-inch-diameter back-projection coated dome centred over the fly platform, which covers 360° in azimuth and 120° in elevation of the fly’s visual field. To simulate an object approaching with constant velocity, the projected looming disc centre remained constant while the disc radius increased nonlinearly over time on the basis of the following equation

$$\theta \left(t\right)=2{{\rm{\tan }}}^{-1}\frac{l}{{vt}}$$

in which θ is the angular size of the stimulus (in radians), l is the radius of the virtual object, and v is its simulated approach velocity. $t = 0$ is the theoretical time of contact, when the object would reach 180°, so that $t < 0$ during object expansion. For optogenetic stimulation, CsChrimson was activated in flies raised on retinal food with four 624-nm wavelength light-emitting diodes (total irradiance of 500 W m^{-2}, as measured from the location of the fly on the platform). Escape responses were captured using a macro lens on a high-speed camera, and two perspectives of the fly (side and bottom views) were filmed at 6,000 frames per second under 850-nm infrared illumination. Only one stimulus was presented per fly, and the platform was cleared before release of the subsequent fly. All looming experiments were carried out during the 4-h activity peak in the afternoon light cycle, and all optogenetic experiments were carried out in the dark.

Behavioural data analysis

Escape sequence durations in the CsChrimson activation and Kir2.1-silencing experiments were manually annotated by labelling the first frame of wing raising and the last frame of tarsal contact from the FlyPEZ video data. For the analysis of postural shifts and takeoff angles following either optogenetic activation or looming stimulus presentation, we used a machine learning software package, Animal Part Tracker (APT, a software package developed by the Branson Lab at Janelia) v0.3.4, which allowed us to automatically track locations of body parts in the input videos. For automated tracking, the videos were subsampled at 600 Hz (1.67-ms interval), which was sufficient to observe smooth changes in leg and body movements. Missing tracking data due to occlusions (body part out of frame) were interpolated for gaps less than five frames (8.33 ms), and a moving-average filter was applied to smooth the raw tracking data. For optogenetic activation experiments, videos in which visibility of T2 legs was lost over the 100 ms of annotation were excluded, except for cases in which the fly performed a takeoff. For silencing experiments, videos in which visibility of T2 legs was lost between the stimulus start and the start of jumping leg extension were excluded from the COM movement, COM flow field and T2 leg angle analyses. Individual takeoff vectors were obtained from two locations of the COM, one at takeoff, when the last of the middle tarsi loses contact with the ground (t_{end}), and one either at a manually

annotated frame of the start of jumping leg extension, or at 5 ms before the takeoff (t_{start}; Fig. 1i). The population mean resultant length, $\bar{R}$, is calculated by the following equation

$$\bar{R}=\frac{1}{n}\left|\mathop{\sum }\limits_{j=0}^{n}{{\rm{e}}}^{i\theta j}\right|$$

in which n is the total number of the takeoff vectors, and ${{\rm{e}}}^{i\theta }$ is Euler's formula as a simplified representation of a vector. $\bar{R}$ is a statistic between 0 and 1 for the spread of a circular variable in the population, such that 1 means all of the takeoff directions are concentrated at a single angle, and 0 means the spread is more uniform. The COM referenced to fly body-centric coordinates was obtained by translating and rotating the COM as described in Extended Data Fig. 1c. Δ[T2 leg angle] at a given time frame of the FlyPEZ video was obtained using the APT-tracked tarsal tips of the middle legs and the COM as described in Fig. 1d. A Butterworth filter was applied to the T2 leg angle time series results. Individual COM movement vectors were calculated as the vector from COM_0 to COM_{pre} (Extended Data Fig. 1d).

Electrophysiological experiments

Electrophysiological recordings and data analysis

Female flies of 2–4 days in age were anaesthetized on a Peltier-driven cold plate and positioned ventral side up to be tethered on a custom polyether-ether-ketone recording plate by applying ultraviolet-cure glue to the head and thorax. We used only female flies because: female flies are larger and hence less prone to desiccation than male flies, and so have the potential to provide longer-lasting electrophysiological recordings; and both the hemibrain and full brain (FAFB) EM datasets were collected from female flies, so our direct measurements of the gradients are both in female flies. For recording stability, the proboscis was glued in a retracted position and the front pair of legs were clipped and glued at the femur. To access the DN soma for whole-cell recording, a window was cut in the cuticle on the posterior side of the head, and the overlying fat and trachea were removed. The brain was continuously perfused during electrophysiology with the

external solution containing (in mM): 103 NaCl, 3 KCl, 5 *N*-Tris (hydroxymethyl)methyl-2-aminoethane-sulfonic acid, 8 trehalose, 10 glucose, 26 $NaHCO_3$, 1 NaH_2PO_4, 1.5 $CaCl_2$ and 4 $MgCl_2$, bubbled with 95% O_2 and 5% CO_2, and adjusted to pH 7.3 and 273–276 mOsm. To disrupt the perineural sheath around the soma of interest, collagenase (0.25 mg ml^{-1} in external solution) was applied locally with a large-bore pipette to the surface of the brain. A small amount of tissue was then removed by using suction from a pipette filled with external solution to gain unrestricted patch pipette access. Patch pipettes were made from borosilicate glass using a Sutter p-1000 puller and fire-polished after pulling using a Narishige MF-900 microforge to achieve a final resistance of 4–8 MΩ. The internal solution contained (in mM): 140 potassium aspartate, 10 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid, 1 ethylene glycol tetraacetic acid, 4 MgATP, 0.5 Na_3GTP and 1 KCl. The pH was 7.3 and the osmolarity was adjusted to approximately 265 mOsm. To obtain patch-clamp recordings, DN somata were visually targeted through brief GFP excitation. Recordings were acquired in current-clamp mode with a MultiClamp 700B amplifier (Molecular Devices), low-pass filtered at 10 kHz, and digitized at 40 kHz (Digidata 1440A, Molecular Devices).

Whole-cell recording data were analysed in MATLAB using custom written code or using Clampfit 11 software (Molecular Devices), and graphical representation was carried out by using Prism 9.2.0 software (GraphPad). Spike events in response to looming stimuli were determined on the basis of the rise slope (mV ms^{-1}) in the response region above a threshold given from the averaged maximum slope in the baseline region across individual recordings, followed by visual inspection of the raw data. The baseline region of each trial corresponded to the 2-s time window before the beginning of the looming stimulus. The response region was the 150-ms period after the onset of the stimulus. To estimate the magnitude of depolarization in response to looming stimuli, membrane potentials were averaged across individual trials (4–8 trials per neuron), and the area (ms × mV) was calculated in the 150-ms response region.

Visual stimulation for electrophysiology

Custom visual stimuli were produced in MATLAB using the Psychophysics Toolbox to display looming stimuli with different approach angles around the fly. We were limited in how far posterior we could show stimuli owing to constraints of the plate to which the fly was tethered to for accessing the back of the head capsule and the microscope. This was especially an issue for DNp11 recordings, as the microscope objective blocks presentation of the posterior stimuli that should most strongly excite DNp11. Thus, our strategy for assessing the functional gradient of the receptive field (RF) was to compare directly measured visual responses in the experimentally accessible visual field to responses predicted by a model we generated from the measured synaptic numbers and an alignment with the visual world (see the section below entitled Mapping the LC4 anatomical RF). Within our accessible visual area, we generated looming stimuli at 32.5°, 45°, 57.5° and 70° along the eye equator (anterior to posterior) and then pitched the plane of these stimuli down 20° to roughly coincide with the tilt of the synaptic gradients we measured. Looming stimuli from different azimuths were shown in randomized sets. Looming stimuli were arrays of three discs, black on a white background, and programmed to expand from 0° to 30° in azimuth in each disc with a 12-s inter-stimulus interval. We used three-disc vertical arrays because we wanted to use a stimulus that would produce as strong a response as possible and which could be varied in azimuth. As LC4 neurons have only an approximately 40° RF, only a handful of LC4 neurons may be excited by a single looming stimulus. Therefore, to activate more LC4 neurons along a given azimuth, we used a column of three. See Extended Data Fig. 4a for a depiction of the looming stimuli used. Visual stimuli were back-projected at 360 Hz onto a 4-inch diameter dome at 768 × 768 resolution. Stimulus frames were synchronized by simultaneously recording a photodiode with the recording trace that monitored a patch of each frame projected just outside the dome and coloured black or white on alternate frames. Constant angular velocity stimuli were generated using the following equation

$$\theta \left(t\right)={v}_{{\rm{a}}}t$$

in which θ is the angular size of the stimulus, ${v}_{{\rm{a}}}$ is the angular velocity, and $\theta = 0$ at $t = 0$. All stimuli were corrected for distortion and irradiance differences as described previously.

$P2X_2$ experiments

Whole-cell patch-clamp recordings from the GF were carried out in 2–4-day-old female flies as described above. For $P2X_2$ receptor activation of LC4 or LPLC2 VPNs, a glass capillary pulled to a 1-µm diameter was positioned on the VPN dendrites, which expressed both GFP and the $P2X_2$ receptor, approximately 50 µm below the surface of the brain. ATP (Sigma A9187, 5 mM) was microinjected (5 psi, 200-ms pulse) under the control of a Picospritzer (Parker Hannifin). To test dorsoventral gradients of functional connectivity between the VPNs and the GF, either the dorsal or ventral part of the lobula was stimulated in an alternating fashion at 90-s intervals to permit recovery between pulses. Whole-cell recording data were analysed as mentioned above. Before calculating the peak amplitudes of the GF response, the membrane potential traces acquired during ATP applications were low-pass filtered and averaged across individual trials as specified in the figure legends.

Generation of single-cell STaR transgenic flies

A combination of HIFI DNA assembly (NEB) and restriction-enzyme-based cloning was used to generate either 13XLexAoP2-FRT-STOP-FRT-myr::GFP-2A-R::PEST or 13XLexAoP2-FRT-STOP-FRT-myr::tdTomato-2A-R::PEST through modification of pJFRC177 (Addgene: 10XUAS-FRT-STOP-FRT-myrGFP, plasmid no. 32149). First, the 10XUAS sequence of pJFRC177 was replaced by 13XLexAoP2 from pJFRC19 (Addgene: 13XLexAoP2-IVS-myrGFP, plasmid no. 26224). Second, the GFP-coding sequence of pJFRC177 was replaced by either GFP-2A (cassette C: GS linker-FRT-STOP-FRT-GFP-2A-LexAVP16) or tdTomato-2A (UAS-DIPalpha-2A-tdTomato), both followed by the coding sequence of R::PEST recombinase from pJFRC165 (Addgene: 20XUAS-IVS-R::PEST plasmid no. 32142). Transgenic flies were generated by integration of either construct into the VK00033 landing site using a commercial injection service (BestGene). To generate sparsely labelled VPNs with visualized presynaptic sites (sparse StaR), 13XLexAoP2-FRT-STOP-FRT-myr::GFP-2A-R::PEST constructs were recombined with StaR[41] (Brp-RSRT-stop-RSRT-myr::smGdP-V5-2A-LexA, laboratory stock). Female flies carrying the recombined constructs were crossed into male flies with VPN-specific

LexA driver lines and hsFLP recombinase. At 48 h after puparium formation, pupae were heat-shocked for 15 min in 37 °C water bath.

Immunohistochemistry

Unless otherwise specified, dissected flies were aged 3–4 days post eclosion. Brains were dissected in ice-cold Schneider's *Drosophila* Medium (Gibco 21720–024), and fixed in acid-free glyoxal (Addax Biosciences) containing 5% sucrose (Sigma S9378) overnight at 4 °C. Brains were rinsed repeatedly with PBST (PBS (Bioland Scientific LLC PBS01-03) containing 0.5% Triton-X100 (Sigma T9284)), and incubated in blocking solution (PBST containing 10% normal goat serum (Sigma G6767)) for 2 h at room temperature before incubation with antibodies. Brains were incubated sequentially with primary and secondary antibodies diluted in blocking solution for 24 h at 4 °C, with three rinses in PBST followed by 1 h incubations at room temperature in between and afterwards. Primary antibodies were used at 1:20 (nc82), 1:500 (chicken anti-GFP) and 1:200 (all others) dilutions. All secondary antibodies were used at 1:300 dilutions. The full list of antibodies used is available in the Reporting Summary. The technique for subsequent mounting in DPX was adapted from the Janelia protocol for mounting the central nervous system of adult *Drosophila* in DPX. After being washed to remove residual secondary antibodies, brains were additionally fixed with PBS containing 4% paraformaldehyde (Electron Microscopy Sciences 15710) for 3 h at room temperature, rinsed with PBS and mounted on 22 × 22-mm square No. 1.5H cover glass (Thorlabs CG15CH2) (with the posterior side of the brain facing the cover glass) previously coated with poly-l-lysine (0.078% solution in deionized water, Sigma P1524) with added 0.2% Kodak Photo-Flo 200 Solution (Electron Microscopy Sciences 74257) followed by a quick 1–2-s rinse with MilliQ water. Brains were dehydrated by placing the cover glass into baths with successively increasing ethanol (Sigma 459844) concentrations (30–50–75–95–100–100–100%, 10 min each) followed by three successive baths of xylene (Thermo Fisher Scientific X5–500), 5 min each. Afterwards the glass was uniformly covered with 8–10 drops of DPX (Electron Microscopy Sciences 13510) and placed on a prepared slide between the spacers made of two 22 × 22 mm square No. 2 cover glasses (Fisher Scientific 12-540B). The

slide was left for 24 h in the hood for drying, and then transferred to room temperature and imaged at least 24 h afterwards,

Confocal image acquisition and processing

Immunofluorescence images were acquired using a Zeiss LSM 880 confocal microscope with Zen digital imaging software using an oil-immersion ×63 objective. Serial optical sections were obtained from whole-mount brains with a typical resolution of 1,024 µm × 1,024 µm × 0.5 µm. Image stacks were exported to Imaris 9.7 for level adjustment, cropping and removal of signal in off-target brain regions and background noise, as well as 3D volume reconstructions.

Analysis of neuroanatomical data from confocal image stacks

To assess and measure the differential placement of DN dendrites within the LC4 glomerulus, confocal image stacks of colocalized glomeruli and DN dendrites were aligned so that the x axis corresponded to the sagittal diameter (width) of the glomerulus and cropped at the edges of the glomerulus to exclude any extraglomerular DN dendrites from consideration. 3D reconstructions of LC4 axon terminals and DN dendrites were obtained using the Imaris Filaments tool (Extended Data Fig. 10b). The x coordinates of the filaments were exported to GraphPad Prism 9.2.0 and normalized to the sagittal diameter of the LC4 glomerulus (0–1 range). The x coordinate of the centroid of the DN dendritic arbour was calculated as a mean of x coordinates of all filaments and used as a final metric of spatial distribution of dendrites within the glomerulus (Extended Data Fig. 10c).

To assess the spatial proximity between presynaptic sites of individual LC4 or LPLC2 neurons and DN dendrites (single-cell STaR experiments), Brp puncta in single VPN cells were reconstructed using the Imaris Spots tool, followed by identification of their centroids, as well as centroids of reconstructed dendritic filaments. Distance between Brp puncta and DN dendrite centroids was measured along the sagittal diameter of the glomerulus (LC4) or along three cardinal axes (A–P, D–V and L–M) of the glomerulus (for LPLC2). Only female flies were used for analysis to be consistent with the available connectome data, which are in a female fly.

Analyser was not blinded to genotype due to characteristic identifiable morphology of DNp02, DNp11 and DNp04, as well as clear anatomical positions of anterior–posterior LC4 and LPLC2.

Connectomics analysis

FAFB connectome reconstruction analysis

We annotated the FAFB serial section transmission EM volume using the CATMAID software to determine the chemical synaptic connectivity between the LC4 neurons and four DNs of interest, DNp02, DNp11, GF and DNp04. As a starting point, we used previously traced skeletons for LC4 neurons. To start tracing the DNs, we used morphological cues from confocal fluorescence imaging in distinct strategies to locate a starting point for tracing each DN. For DNp02, confocal microscopy stacks suggested that the somata neurite travels close to the path of the GF somata neurite. We found DNp02 by locating its neurite within a shared soma tract, which, along with several other neurites, appears encased in a dark sheath. DNp04 was located when tracing the LC4 neurons. The skeleton was then traced out and linked to the same soma tract as DNp02 and GF. DNp11 was located by searching for candidate DNs that cross the midline dorsal of the oesophagus. From each starting node, the full skeleton was traced and compared to the confocal image stacks for confirmation of cell type identity. To determine the chemical synaptic connectivity, we searched for four criteria: T-bars, presynaptic vesicles, synaptic clefts and postsynaptic densities. If a potential synapse possessed two out of four criteria, it was labelled as a synapse. We focused our efforts on LC4 (presynaptic) and DNp02, DNp11, GF and DNp04 (postsynaptic) synapses to gain a representative view of the connectivity between LC4 and the DNs.

Mapping the LC4 anatomical RF

To model the real-world RFs of the LC4 population, we followed a previously established method[25], and applied it to newly reconstructed LC4 neurons. We first mapped all 55 LC4 dendrites (FAFB volume) onto a layer of the lobula by fitting a second-order surface to all of the dendritic arbours. Each projected dendrite traced out a polygon that represented the field of

view of the corresponding LC4 neuron. We modelled each LC4 as a 2D circular Gaussian on this surface. Its height was set to be unity, and its width was given by the radius of a circle that had the same area as the projected polygon. To map each LC4 neuron's location (COM of the dendrite) onto eye coordinates, we used as reference points previously reconstructed Tm5 neurons[25] from two medulla columns, which correspond to the centre of the eye and a dorsal position on the central meridian (the line that partitions the eye between anterior and posterior halves). To estimate an LC4-DN's RF, we first multiplied each LC4 Gaussian's height by the number of synaptic connections to that LC4-DN. We then summed all LC4 Gaussians to produce a 2D multi-Gaussian distribution, which was the LC4-DN's RF. To estimate an LC4-DN's response to a looming stimulus, we multiplied the LC4 Gaussian's height by both the number of synaptic connections and the percentage of the LC4 RF that was covered by the stimulus at its maximum size (30°). For instance, if the stimulus overlapped with 40% of an LC4's RF, then that LC4 Gaussian's effective height was the number of connections times 0.4. Finally, all LC4 contributions were summed to produce the estimated response of the LC4-DN to the looming stimulus. Note that LC4s that did not overlap at all with a stimulus contributed nothing to the DN's response.

Hemibrain connectome reconstruction analysis

Volumetric data of neurons and neuropils, as well as connectivity data and synapse locations, were obtained from the neuPrint (hemibrain v1.1) database, (https://neuprint.janelia.org/) and have been processed with the natverse package[51] for R (v4.0.3) using custom scripts. All coordinates in these datasets are based on the original voxel resolution of 8 nm.

k-means clustering of individual neurons within VPN cell type populations

For each VPN cell type, a matrix of synaptic connections between individual VPN neurons and their postsynaptic partners was constructed using the neuprintR package. Postsynaptic partners forming fewer than 50 total synapses with the entire VPN cell type population were excluded (about 1 synapse per individual VPN on average; we reasoned that this threshold

would reflect the limit of EM data reconstruction error rate). Synaptic connections within the population of VPN cell type were also removed (for example, LC4 to LC4 synapses). The resulting matrix was scaled such that the variables (individual postsynaptic partners) had unit variance across the observations (individual VPN cells in the population). Principal component analysis was carried out on the scaled matrix. Up to ten principal components were used for *k*-means clustering on the individual VPNs (the number of PCs was determined on the basis of the drop in the eigenvalues in the scree plots for each VPN type). A value of *k* was subsequently determined from the corresponding scree plots by the drop in the within-cluster sum of squared distance (example in Extended Data Fig. 6a).

Correlation in synaptic connectivity

Matrices of correlation in synaptic connectivity (Fig. 4d,f) were generated using the pairwise Spearman's correlation coefficient of the 300 unique pairs derived from the top 25 postsynaptic partners (based on the total number of synapses and excluding connections with the same VPN cell type) of LC4 and LPLC2, ordered using hierarchical clustering. Each entry evaluates the monotonic relationship between a pair of the synaptic connectivity gradients. For each pair, the correlation coefficient was calculated using the vectors containing the number of synapses between the selected postsynaptic partners and each individual VPN cell within the population (example in Fig. 2h).

Weighted cendritic centroids

To evaluate the distances between weighted dendritic map centroids for each postsynaptic partner of LC4 and LPLC2, we identified the endpoints of the dendrites innervating the lobula for each individual VPN cell. These were isolated using cut-planes that were manually selected to optimally separate the lobula region (Extended Data Fig. 9a–d). We then evaluated the centroid of the selected endpoints by calculating their spatial average. We repeated these steps for all VPN cells within a population (71 for LC4 and 85 for LPLC2). The resulting 3D centroids were then projected onto the cut-plane. The outlines of the lobula were obtained by evaluating the convex hull of the projections of all the selected endpoints for all of the cells of the examined

VPN. To identify a weighted innervation centroid for a given postsynaptic partner, we calculated the overall weighted median using the number of synapses associated with each centroid as weights. We then identified the top anticorrelated pairs of postsynaptic partners by selecting those for which the Spearman's correlation coefficient is below a certain threshold that was determined by evaluating, for each VPN, the value that optimizes the correlation between the dendritic map and the synaptic connectivity correlation. For each one of these top pairs, we estimated the perpendicular to the line connecting the corresponding weighted median centroids. These lines were combined using the median operator to reduce the influence of potential outliers. This resulted in a single line identifying the optimal unbiased separator of the most anticorrelated pairs (median separation line in Fig. 4e,g). The distance between their projections onto the line perpendicular to the optimal separator (projection line in Fig. 4e,g) was used as a final metric to generate the matrix and calculated for each pair of postsynaptic partners (Extended Data Fig. 9g,h). The projection line for LC4 was almost parallel to the A–P axis of the lobula (Fig. 4e), and slightly deviated from that for LPLC2 owing to the dual nature of synaptic gradients in this cell type (both A–P and D–V).

Spatial distribution of postsynaptic sites in optic glomeruli

A similar approach based on the estimation of an unbiased separator was used to evaluate the correlation between the centroids of postsynaptic sites for postsynaptic partners of VPNs. To estimate this separator, we started by isolating all postsynaptic sites within the glomerulus using a cut-plane. We then selected the top anticorrelated pairs of postsynaptic partners, in a manner similar to how we analysed the dendritic map centroids. For each pair, we split the postsynaptic sites into two different classes depending on the postsynaptic partner they belonged to and used a support vector machine with a linear kernel to evaluate the optimal separating plane. We then computed the median of these planes. This resulted in a single plane identifying the unbiased optimal separator of the most anticorrelated pairs (median separation plane in Fig. 6b,j). We then projected the postsynaptic sites of each postsynaptic partner onto the line perpendicular to the optimal separator and calculated the distance between the median of the respective projections. The distance matrices for a given VPN cell type were obtained

by calculating the pairwise distances between each of the 300 pairs of postsynaptic partners of LC4 and LPLC2 (Extended Data Fig. 9i,j). For selected pairs of postsynaptic neurons, the distributions of postsynaptic sites projections were compared using the two-sample Kolmogorov–Smirnov test (Fig. 6b,j).

Assessment of topographic mapping in VPN optic glomeruli

Skeletons of individual neurons within each VPN cell type were selected manually on the basis of A–P and D–V topographic location of their dendrites and/or the pattern of *k*-means clustering of the dendritic maps (15 cells per topographic domain, unless stated otherwise in the figure legends). Groups of neurons with dendrites in different topographic domains were differentially coloured. Axonal processes of the corresponding neurons were traced in the optic glomerulus and visually examined for traces of spatially ordered organization. LC10 neurons were excluded from the analysis owing to previously reported A–P axonal topography[6,13]. LC6 neurons were excluded owing to previous extensive analysis[25] indicating the presence of coarse glomerular retinotopy inaccessible through visual examination.

Statistical analysis

All statistical analyses were carried out in RStudio 1.4.1103, MATLAB or Prism 9.2.0 software (GraphPad). NS: $P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$ for all figures where applicable. Statistical tests for Figs. 1e and 3e,h and Extended Data Figs. 1,2,4 and 12 are described in Supplementary Table 1. In all box plots (Fig. 6 and Extended Data Fig. 11), the solid line depicts the median; the upper and lower bounds of the box depict the third and first quantiles of the data spread, respectively. Whiskers indicate minimum and maximum values. All other statistical tests, number of replicates, statistical significance levels and other elements of statistical analysis are reported in the corresponding section of the Methods, along with the associated results and/or in the corresponding figure legends. No data were excluded from the analysis except as noted for the behaviour experiments (see the section in the Methods entitled Behavioural data analysis). All measurements were taken from distinct samples.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets generated during the current study are available as downloadable files at https://data.mendeley.com/datasets/pnbmx825wv, https://data.mendeley.com/datasets/84kh3ncbf8 and https://data.mendeley.com/datasets/th99hk824v. These include confocal image stacks related to Figs. 4–6. Other datasets generated and/or analysed during the current study are available from the corresponding authors on reasonable request. For further information regarding any resources and reagents, contact G.M.C. (gwyneth.card@columbia.edu) or S.L.Z. (lzipursky@mednet.ucla.edu). Source data are provided with this paper.

Code availability

R scripts for hemibrain connectome reconstruction analysis are available at https://github.com/avaccari/DrosophilaVPNWiring. Code associated with APT v0.3.4 (developed by the Branson Lab at Janelia) is available at https://zenodo.org/record/6366082. All reconstructed neurons from the FAFB dataset will be made available at https://fafb.catmaid.virtualflybrain.org/. All code to carry out the analysis and generate the figures on the FAFB data will be available at https://github.com/artxz/LC4_code.

References

1. Crochet, S., Lee, S. H. & Petersen, C. C. H. Neural circuits for goal-directed sensorimotor transformations. *Trends Neurosci.* **42**, 66–77 (2019).

2. Calhoun, A. J. & Murthy, M. Quantifying behavior to solve sensorimotor transformations: advances from worms and flies. *Curr.*

Opin. Neurobiol. **46**, 90–98 (2017).

3. Cavanaugh, J. et al. Optogenetic inactivation modifies monkey visuomotor behavior. *Neuron* **76**, 901–907 (2012).

4. Bianco, I. H. & Engert, F. Visuomotor transformations underlying hunting behavior in zebrafish. *Curr. Biol.* **25**, 831–846 (2015).

5. Buneo, C. A., Jarvis, M. R., Batista, A. P. & Andersen, R. A. Direct visuomotor transformations for reaching. *Nature* **416**, 632–636 (2002).

6. Wu, M. et al. Visual projection neurons in the *Drosophila* lobula link feature detection to distinct behavioral programs. *Elife* **5**, e21022 (2016).

7. Otsuna, H. & Ito, K. Systematic analysis of the visual projection neurons of *Drosophila melanogaster*. I. Lobula-specific pathways. *J. Comp. Neurol.* **497**, 928–958 (2006).

8. Coen, P., Xie, M., Clemens, J. & Murthy, M. Sensorimotor transformations underlying variability in song intensity during *Drosophila* courtship. *Neuron* **89**, 629–644 (2016).

9. Huston, S. J. & Jayaraman, V. Studying sensorimotor integration in insects. *Curr. Opin. Neurobiol.* **21**, 527–534 (2011).

10. Helmbrecht, T. O. et al. Topography of a visuomotor transformation. *Neuron* **100**, 1429–1445 (2018).

11. Heukamp, A. S., Warwick, R. A. & Rivlin-Etzion, M. Topographic variations in retinal encoding of visual space. *Annu. Rev. Vis. Sci.* **6**, 237–259 (2020).

12. Klier, E. M., Wang, H. & Crawford, J. D. The superior colliculus encodes gaze commands in retinal coordinates. *Nat. Neurosci.* **4**, 627–632 (2001).

13. Timaeus, L., Geid, L., Sancer, G., Wernet, M. F. & Hummel, T. Parallel visual pathways with topographic versus nontopographic organization

connect the *Drosophila* eyes to the central brain. *iScience* **23**, 101590 (2020).

14. Aptekar, J. W., Keleş, M. F., Lu, P. M., Zolotova, N. M. & Frye, M. A. Neurons forming optic glomeruli compute figure–ground discriminations in *Drosophila*. *J. Neurosci.* **35**, 7587–7599 (2015).

15. Keleş, M. F. & Frye, M. A. Object-detecting neurons in *Drosophila*. *Curr. Biol.* **27**, 680–687 (2017).

16. Klapoetke, N. C. et al. A functionally ordered visual feature map in the *Drosophila* brain. *Neuron* **110**, 1700–1711 (2022).

17. Städele, C., Keleş, M. F., Mongeau, J. M. & Frye, M. A. Non-canonical receptive field properties and neuromodulation of feature-detecting neurons in flies. *Curr. Biol.* **30**, 2508–2519 (2020).

18. Bidaye, S. S. et al. Two brain pathways initiate distinct forward walking programs in *Drosophila*. *Neuron* **108**, 469–485 (2020).

19. Ribeiro, I. M. A. et al. Visual projection neurons mediating directed courtship in *Drosophila*. *Cell* **174**, 607–621 (2018).

20. Sen, R. et al. Moonwalker descending neurons mediate visually evoked retreat in *Drosophila*. *Curr. Biol.* **27**, 766–771 (2017).

21. Zacarias, R., Namiki, S., Card, G. M., Vasconcelos, M. L. & Moita, M. A. Speed dependent descending control of freezing behavior in *Drosophila melanogaster*. *Nat. Commun.* **9**, 3697 (2018).

22. Ache, J. M., Namiki, S., Lee, A., Branson, K. & Card, G. M. State-dependent decoupling of sensory and motor circuits underlies behavioral flexibility in *Drosophila*. *Nat. Neurosci.* **22**, 1132–1139 (2019).

23. Cande, J. et al. Optogenetic dissection of descending behavioral control in *Drosophila*. *Elife* **7**, e34275 (2018).

24. Namiki, S., Dickinson, M. H., Wong, A. M., Korff, W. & Card, G. M. The functional organization of descending sensory-motor pathways in *Drosophila*. *Elife* **7**, e34272 (2018).

25. Morimoto, M. M. et al. Spatial readout of visual looming in the central brain of *Drosophila*. *Elife* **9**, e57685 (2020).

26. Panser, K. et al. Automatic segmentation of *Drosophila* neural compartments using GAL4 expression data reveals novel visual pathways. *Curr. Biol.* **26**, 1943–1954 (2016).

27. Scheffer, L. K. et al. A connectome and analysis of the adult *Drosophila* central brain. *Elife* **9**, e57443 (2020).

28. Card, G. & Dickinson, M. H. Visually mediated motor planning in the escape response of *Drosophila*. *Curr. Biol.* **18**, 1300–1307 (2008).

29. Tanaka, R. & Clark, D. A. Neural mechanisms to exploit positional geometry for collision avoidance. *Curr. Biol.* **32**, 2357–2374 (2022).

30. Muijres, F. T., Elzinga, M. J., Melis, J. M. & Dickinson, M. H. Flies evade looming targets by executing rapid visually directed banked turns. *Science* **344**, 172–177 (2014).

31. Fotowat, H. & Gabbiani, F. Collision detection as a model for sensory-motor integration. *Annu. Rev. Neurosci.* **34**, 1–19 (2011).

32. Eaton, R. *Neural Mechanisms of Startle Behavior* (Springer, 1984).

33. Williamson, R., Peek, M. Y., Breads, P., Coop, B. & Card, G. M. Tools for rapid high-resolution behavioral phenotyping of automatically isolated *Drosophila*. *Cell Rep.* **25**, 1636–1649 (2018).

34. Ache, J. M. et al. Neural basis for looming size and velocity encoding in the *Drosophila* giant fiber escape pathway. *Curr. Biol.* **29**, 1073–1081 (2019).

35. von Reyn, C. R. et al. Feature integration drives probabilistic behavior in the *Drosophila* escape response. *Neuron* **94**, 1190–1204 (2017).

36. Von Reyn, C. R. et al. A spike-timing mechanism for action selection. *Nat. Neurosci.* **17**, 962–970 (2014).

37. Card, G. M. Escape behaviors in insects. *Curr. Opin. Neurobiol.* **22**, 180–186 (2012).

38. Zheng, Z. et al. A complete electron microscopy volume of the brain of adult *Drosophila melanogaster*. *Cell* **174**, 730–743 (2018).

39. Klapoetke, N. C. et al. Ultra-selective looming detection from radial motion opponency. *Nature* **551**, 237–241 (2017).

40. Nern, A., Pfeiffer, B. D. & Rubin, G. M. Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. *Proc. Natl Acad. Sci. USA* **112**, E2967–E2976 (2015).

41. Chen, Y. et al. Cell-type-specific labeling of synapses in vivo through synaptic tagging with recombination. *Neuron* **81**, 280–293 (2014).

42. Georgopoulos, A. P., Kettner, R. E. & Schwartzb, A. B. Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *J. Neurosci.* **8**, 2928–2937 (1988).

43. Namiki, S. et al. A population of descending neurons that regulates the flight motor of *Drosophila*. *Curr. Biol.* **32**, 1189–1196 (2022).

44. Hamon, A., Guillet, J. C. & Callec, J. J. A gradient of synaptic efficacy and its presynaptic basis in the cercal system of the cockroach. *J. Comp. Physiol. A.* **167**, 363–376 (1990).

45. Fisher, Y. E., Lu, J., D'Alessandro, I. & Wilson, R. I. Sensorimotor experience remaps visual input to a heading-direction network. *Nature* **576**, 121–125 (2019).

46. Pinto-Teixeira, F. et al. Development of concurrent retinotopic maps in the fly motion detection circuit. *Cell* **173**, 485–498 (2018).

47. Flanagan, J. G. Neural map specification by gradients. *Curr. Opin. Neurobiol.* **16**, 59–66 (2006).

48. Kurmangaliyev, Y. Z., Yoo, J., Valdes-Aleman, J., Sanfilippo, P. & Zipursky, S. L. Transcriptional programs of circuit assembly in the *Drosophila* visual system. *Neuron* **108**, 1045–1057 (2020).

49. Cheng, S. et al. Vision-dependent specification of cell types and function in the developing cortex. *Cell* **185**, 311–327 (2022).

50. Shinomiya, K. et al. The organization of the second optic chiasm of the *Drosophila* optic lobe. *Front. Neural Circuits* **13**, 65 (2019).

51. Bates, A. S. et al. The natverse, a versatile toolbox for combining and analysing neuroanatomical data. *Elife* **9**, e53350 (2020).

Acknowledgements

We thank the Bloomington Stock Center and Janelia Fly Facility for providing fly stocks, and we thank the Janelia Fly Facility for providing genetic crosses for the behaviour screen. We thank members of the Janelia Connectome Annotation Team (now Annotation & Analytics), including M. Ryan and S. Murthy, for manual reconstruction of neurons in the FAFB dataset, and R. Parekh for CAT team management. We thank M. Dickinson for use of his flycon art in Fig. 1 and Extended Data Figs. 1 and 2. We thank all members of the Zipursky Lab, T. Oram, S. Asinof and members of the Card Lab for providing advice and comments on the manuscript. This work was supported by the Helen Hay Whitney Foundation (M.D). S.L.Z. and G.M.C. are investigators of the Howard Hughes Medical Institute.

Author information

Author notes

1. Shigehiro Namiki

Present address: Research Center for Advanced Science and Technology, University of Tokyo, Tokyo, Japan

2. Gwyneth M. Card

Present address: Department of Neuroscience, Howard Hughes Medical Institute, The Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA

3. These authors contributed equally: Mark Dombrovski, Martin Y. Peek, Jin-Yong Park

Authors and Affiliations

1. Department of Biological Chemistry, Howard Hughes Medical Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

   Mark Dombrovski, Yerbol Z. Kurmangaliyev, Piero Sanfilippo, Aadil Rehan & S. Lawrence Zipursky

2. Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA

   Martin Y. Peek, Jin-Yong Park, Marissa Sumathipala, Carmen Morrow, Patrick Breads, Arthur Zhao, Jason Polsky, Shada Alghailani, Emily Tenshaw, Shigehiro Namiki & Gwyneth M. Card

3. Department of Computer Science, Middlebury College, Middlebury, VT, USA

   Andrea Vaccari

Contributions

G.M.C., M.Y.P., M.D. and S.L.Z. conceived the project and designed the research; M.Y.P., J.-Y.P., M.D., A.R., J.P., S.A. and E.T. carried out experiments and collected data; J.-Y.P., M.D., A.V., M.S., A.Z., C.M., P.B.,

Y.Z.K. and S.N. analysed data; A.V., Y.Z.K., P.S., A.Z. and S.N. generated resources; M.D., M.Y.P., J.-Y.P., G.M.C. and S.L.Z. wrote, reviewed and edited the paper; G.M.C. and S.L.Z. acquired financial support for the project

Corresponding authors

Correspondence to S. Lawrence Zipursky or Gwyneth M. Card.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Hokto Kazama, Daniel Tomsic and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Control of fly takeoff direction.

a, When shown a looming stimulus from the front (azimuth 0°), side (azimuth 90°), or back (azimuth 180°), flies respond by moving their center of mass (COM) away from the stimulus. Black disc represents stimulus location and color indicates time from stimulus onset. Flies used for **a**–**d**

were a control genotype for other experiments, Empty>Kir2.1. One trial per fly. **b**, Some flies also takeoff in response to looming, and those that do takeoff in a direction away from the stimulus (with some influence of the heading of the fly[33]). Shown are polar takeoff direction histograms with 12° bin width and mean resultant vector overlaid (red line). p, Hodges-Ajne test for angular uniformity. **c**, Takeoff direction results from the fly shifting its COM relative to the axes formed by a line connecting the ground contact points of its two middle jumping legs and a perpendicular bisector. Black points indicate COM at stimulus onset and red points indicate COM just prior to takeoff. **d**, The specific direction in which the COM moves in body coordinates depends on its starting location. Vector position is the COM position at stimulus onset. The vector itself indicates the shift of COM position from stimulus onset to just prior to takeoff. Black vectors are tracked data, gray vectors are interpolated. Black square is approximated point of convergence. **e**, Percent of flies (individual DN driver lines) that performed a takeoff in response to CsChrimson optogenetic activation in the FlyPEZ assay. Error bars, Wilson score interval; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$ vs control (Empty, empty brain split-Gal4 control; DL – wild type control); normal approximation to binomial, two-sided Z-test, Bonferroni correction post hoc test. **f**, Same data as in (**e**) but with driver lines grouped by cell type. Error bars, SD. **g**, Histograms displaying the distribution of escape sequence durations between the wing raising and takeoff jump sub-behaviors (for LC4-DN driver lines expressing CsChrimson that can elicit escape upon activation). Escape trials are combined from split-Gal4 lines for each LC4-DN type. Short-mode escape duration (0 to 7 ms, gray shaded region) and long-mode escape duration (>7 ms), as previously established. **h**, Percentage of short-mode activated escapes. Error bars, Wilson score interval; $****p < 0.0001$ versus GF; normal approximation to binomial, two-sided Z-test, Bonferroni correction post hoc test. Detailed description of statistical tests used and p-values for panels "e" and "h" is available in Supplementary Table 1.

Source data

Extended Data Fig. 2 Silencing of either DNp02 or DNp11 impairs control of postural shifting and takeoff direction in response to looming stimuli.

For neuronal silencing experiments, driver lines for DNp02, DNp11, and an 'Empty' driver line control were crossed into UAS-Kir2.1. **a**–**b**, Polar takeoff direction histograms in response to looming stimuli presented in front of the fly at 0° azimuth (**a**) or behind the fly at 180° azimuth (**b**); 12° bin width; red line, mean resultant vector. DN-silenced flies perform normally in response to a posterior (180°) stimulus compared to control ($p > 0.1$ for both DNs, Kuiper's Test). However, DN silencing altered the distribution of backward takeoffs direction in response to frontal looming (0°) for both DNp02 ($p < 0.005$, Kuiper's test) and DNp11 ($p < 0.001$, Kuiper's test) silencing compared to controls. Strikingly, many DNp02- and DNp11-silenced flies performed forward takeoffs in response to frontal looming stimulation, effectively jumping toward the threatening stimulus. **c**, To further understand why flies were inappropriately taking off forwards, we looked at how much DN-silenced flies moved their COM backwards in response to 0° looming. We visualized COM movement in body coordinates from different starting postures using the same flow fields in body-centric coordinates as in Extended Data Fig. 1d. Visual inspection indicated that COM movement fields for DN-silenced flies differed from controls in the amount of backwards movement and had more lateral movement. **d**, To quantify this motion, we measure the T2 angle (angle formed by T2 tarsal contact points and COM), which is >180° when the COM is in front of the T2 jumping legs and <180° when the COM is behind the T2 jumping legs. The mean T2 angle just before takeoff was significantly different for DNp02- and DNp11-silenced flies compared to controls (*$p = 0.0468$, ***$p = 4.79e-04$, One-Way ANOVA, Dunnett's test). Black points, individual flies; error bars, SD. **e**, Looking at time courses for T2 leg angle in response to 0° azimuth looming stimulus for the different DN-silenced lines (colors, shaded area, SD), with control data overlaid (grey), it is clear that the difference in the DN-silenced flies is that they do not shift backwards as much as controls. Since COM placement prior to takeoff determines whether the fly's jump will propel it forwards (T2 angle>180) or backwards (T2 angle<180), the impaired pre-takeoff T2 leg angle change in DNp02- and DNp11-silenced flies, which on average does not become <180° as in control flies, likely underlies altered takeoff performance leading to more forward-directed takeoffs. **f**, DNp02 and DNp11 silencing does not affect takeoff rates. Percentage of flies which performed a takeoff to a looming visual stimulus (azimuth = 90°, elevation = 45°) at four looming rates (l/v =

10, 20, 40 and 80 ms), or a looming visual stimulus (azimuth = 0° or 180°, elevation = 45°) at l/v = 40. L1/L2-silenced flies serve as "motion-blind" negative controls. Error bars, SEM; Wilson score interval; **p < 0.01, ***p < 0.001, ****p < 0.0001 versus Empty control; normal approximation to binomial, two-sided Z-test, Bonferroni correction post hoc test. Detailed description of statistical tests and p-values for panel "f" is available in Supplementary Table 1.

Source data

Extended Data Fig. 3 EM-based analysis if synaptic connectivity patterns between LC4 and DNs.

a, Individual LC4 neurons (55 total traced in "FAFB" EM volume) make different numbers of synapses with different DNs. Individual LC4 neurons (x-axis) are ordered according to the synapse number with DNp02 (red bars). LC4 synapse number with DNp11 (blue bars) and DNp04 (purple bars). **b**, LC4 neurons (grey) and synaptic partners (GF and DNp04) traced in a FAFB EM volume. DN soma, circles. **c**, LC4 dendrites (lateral views) color-coded according to the total number of synapses onto each DN. D, dorsal; L, lateral; A, anterior.

Source data

Extended Data Fig. 4 In vivo whole-cell recordings from DNs upon looming stimulation.

a, Left: Schematic of looming array stimuli at different azimuths (32.5°, 45°, 57.5° and 70°). Each loom consists of three dark disks in the white background. Pseudo-colored for clarity. Middle: looming array disk size over time from the beginning of stimulus. Each disk expands from 0° to 30° at 500°/s. Right: Stimulus arrangement projected onto fly's eye (0° front of fly). To align looming array stimuli more closely with the synaptic gradients, the whole plane of the stimuli was pitched down 20°. **b**, Top: whole-cell electrophysiological recordings of the GF and DNp04 to looming stimuli at 32.5° in azimuth. Shown are representative traces from a single fly and stimulus. Middle: change of disk size over time. Bottom: baseline region and

response region defined in the traces for analysis of DN activity. **c**, Representative DN responses showing identified spikes (top rasters). **d**, Representative responses from a single fly for 32.5° (top) and 70° (bottom) azimuth looming stimuli. Representative traces for GF and DNp04 are from a single fly with 6 trials to each stimulus. **e**, Spike raster plots of DNp04 in 150 ms time window after the onset of looming stimuli. **f**, Pooled mean of DNp04 spike numbers across individual flies (from "e"). Error bars, SEM; RM one-way ANOVA, Dunnett's test. **g**, Averaged response of the representative traces in **d** shows subthreshold depolarizing responses to looming stimuli. Shaded area under the line shows estimated depolarization from the baseline. **h**, Pooled mean of integrated potentials for DNp04 and GF across individual flies. Error bars, SD; RM one-way ANOVA, Dunnett's test. **i**–**j**, Mean spike numbers (**i**) and mean of integrated potentials (**j**) across trials in individual flies in response to looming stimuli. Colored lines denote the representative traces of each DN in Fig. 3a and Extended Data Fig. 4b. Detailed description of statistical tests and p-values for panels "f, h" is available in Supplementary Table 1.

Source data

Extended Data Fig. 5 EM reconstruction of LC4 neurons and estimated receptive fields.

a, Tracing of all 55 LC4 neurons on one side of the brain in FAFB EM volume. Two example neurons are colored in red and blue (same in "b"). LC4 neurons have dendrites in lobula layers 2 and 4. LC4 cell bodies are marked with grey circles. Two Tm5 neurons are chosen to determine the center and central meridian of the eye in "b". **b**, Two-dimensional projection of the lobula layer. Traced LC4 dendrites (grey) are projected onto a surface fit through all dendrites (orange). Blue circles denote centers of mass of individual LC4 dendrites. Vertical line is the estimated central meridian that divides the eye between anterior and posterior halves. **c**, Pseudo-stimuli mimicking looming stimuli used in in vivo whole-cell recordings to simulate LC4DN responses based on LC4 connectivity. **d**, Confined DN RFs to the region of pseudo-stimuli in the eye coordinate imitating the looming stimuli. LC4 RFs based on connectivity to each DN are scaled by the proportion of overlapping with each stimulus region prior to sum up all the LC4 RFs. **e**,

Estimated anatomical receptive fields (RFs) of the GF and DNp04 in the eye coordinates. **f**, Left: Estimated connectivity-based DN responses to pseudo-stimuli (solid) and measured responses to the corresponding looming stimuli (dashed) from *in vivo* whole-cell recordings. Right: Estimated responses in DNp04 and GF extrapolated using the connectivity model to looming pseudo-stimuli that cover the visual hemifield.

Extended Data Fig. 6 Differential synaptic connectivity in VPNs manifests as synaptic gradients reflecting visual space map.

a, Representative elbow plots from k-means clustering of individual cells within each VPN cell type (based on the number of synapses they make with postsynaptic partners). A drop in the within cluster sum of squared distance was used to determine the number of k in Fig. 4a, b (see Methods for details). **b**, Representative example of differential synaptic connectivity metric (median coefficient of variation) for LC4. Left: coefficient of variation in synapse number between individual neurons within LC4 population and top 25 postsynaptic partners of LC4 (ordered by decreasing variation). Right: summary (median CV, shown in red) metric for all postsynaptic partners of LC4 making >50 synapses total. **c**, Representative examples of graded synaptic connectivity between four VPN cell types and their top 15 postsynaptic partners based on the total number of synapses. Each individual neuron within a VPN cell type is assigned a color based on just one plot (DNp11 for LC4, Giant Fiber for LPLC2 etc.), with the colors preserved in other graphs. Every plot indicates the number of synapses between individual neurons within one VPN cell type and a given postsynaptic partner (arranged by descending number of synapses). **d**, Single LC4 neurons (EM-based connectome reconstructions) with dendrites in anterior (bodyID 1907587934) or posterior (bodyID 1249932198) regions of the lobula are highlighted. The remaining LC4 neurons shown in grey. **e**, Differential synaptic connectivity between two LC4 neurons from (**d**) and their top 25 postsynaptic partners (measured by total number of synapses). 15 out of 25 postsynaptic neurons receive preferential or exclusive input from either anterior or posterior LC4. **f**–**g**, Differential synaptic connectivity of individual LPLC2 neurons with dendrites in dorsal (bodyID 1815826155) vs ventral (bodyID 1815809293) lobula. Similar to (**d**, **e**). P, posterior; M, medial; D, dorsal; L, lateral.

Extended Data Fig. 7 Synaptic gradients are a common wiring motif across 20 VPN cell types.

Additional examples of synaptic gradients reflecting the visual space map between different VPN cell types and their postsynaptic partners. See legend in Fig. 2c,d. Top: 2D lobula projections of dendritic centroids for individual VPNs in the lobula weighted in size and color by the number of synapses made by their axons onto a designated postsynaptic target. Bottom: regression of synaptic weights as a function of VPN dendrite centroid location along the AP or DV axis in the lobula. Linear fit line overlaid. Error bands, SEM. low R^2 value of LC10 gradients may be explained by the fact that LC10 bodyID annotations in the hemibrain EM volume contain multiple LC10 subtypes (LC10a-d) with differential synaptic specificity. D, dorsal; A, anterior.

Extended Data Fig. 8 Topographic mapping in VPN optic glomeruli.

a–**d**, Examples of VPN cell types with retinotopic mapping of axon terminals reflecting the AP axis of the lobula dendritic map. Assessed via EM reconstructions and light microscopy (individual cells are labeled using Multicolor Flp-Out). LC4 neurons (**a**, same pair as in Fig. 5a) with dendrites innervating anterior and posterior lobula have axon terminals in distinct regions of the glomerulus. LC9 (b), LC22 (c), LPLC4 and LPLC1 (d) show similar axonal topography. Left panels, hemibrain connectome reconstructions of 15 anterior (red), 15 posterior (blue), and the remaining cells (grey). Corresponding reconstruction of axons in the VPN glomeruli shows visual separation of anterior and posterior terminals. Right panels: one anterior (red) and one posterior (green) cell labeled using Multicolor Flp-Out and assessed via light microscopy. (n = 4 pairs of A-P individual clones from different brains with reproducible axon terminal topography for LPLC4 and LPLC1, n = 5 for LC9, n = 6 for LC22). **e**, Traces of DV axonal retinotopy in LC16 – a single example found across 20 VPN cell types. **f**, Representative examples of VPNs without topographic mapping of axon terminals in optic glomeruli (EM reconstructions), despite elaborating synaptic gradients reflecting visual space map (see Fig. 4h and Extended Data Fig. 7). **g**–**j**, Differential axon terminal morphology and glomerular

targeting between individual LC4 (**g**, **h**) and LC22 neurons (**i**, **j**) innervating anterior and posterior regions of the lobula (N = 9 pairs of A-P individual clones from different brains with reproducible axon terminal topography for LC4, n =64 for LC22, correspond to examples from Fig. 5). Characteristic branching patterns are consistent between light microscopy and EM-based connectome reconstruction. A, anterior; P, posterior; D, dorsal; V, ventral; L, lateral; M, medial.

Extended Data Fig. 9 EM-based analysis of synaptic connectivity patterns in LC4 and LPLC2.

a–**d**, Analysis of the spatial distribution of dendritic centroids in the lobula for LC4. Centers of mass (centroids) of dendrites for individual 71 LC4 neurons (red dots) were established based on coordinates of end branching points (green dots) located laterally from the separation cut-plane (grey) to exclude the branching points in the glomerulus. Branching points and centroids were projected onto the plane (yellow and blue dots, respectively). Shown for the entire LC4 population and two representative neurons from posterior (**a**–**b**) and lateral (**c**–**d**) views. Similar analysis was also done for LPLC2. D, dorsal; M, medial; A, anterior. **e**–**f**, Relationship between synaptic input specificity and spatial dendritic map (measured by positions of weighted dendritic centroids) for 300 pairs of top 25 postsynaptic partners of LC4 (**e**) and LPLC2 (**f**). r_s, Spearman's rank correlation coefficient. Error bands, SEM. Strong correlation indicates that LC4 and LPLC2 neurons with neighboring dendrites have similar synaptic specificity in the glomerulus. **g**–**h**, Matrices of pairwise distances between weighted centroids of inputs for 300 pairs of the top 25 postsynaptic partners of LC4 (**g**) and LPLC2 (**h**). The order of postsynaptic neurons is preserved from Fig. 4d and f. **i**–**j**, Matrices of pairwise distances between centroids of postsynaptic sites of the top 25 postsynaptic partners of LC4 (**i**) and LPLC2 (**j**). The order of postsynaptic neurons is preserved from Fig. 4d and f.

Source data

Extended Data Fig. 10 Assessment of wiring strategy in VPN glomeruli using light-level neuroanatomy.

a, Top: EM-based connectome reconstructions of LC4 neurons (green) and three DNs. Bottom: confocal projections of colocalized LC4 and three DNs, LC4 glomerulus is indicated with a dashed yellow rectangle (n = 12 brains for each LC4-DN, corresponding to Fig. 5c,d). Note that DNp11 has an additional dendritic branch in the lobula. **b**, Imaris reconstructions of confocal image stacks: LC4 glomerulus (axons) and dendritic segments of three DNs (both reconstructed as filaments) as indicated. **c**, Assessment of spatial distribution of DN dendrites within the volume of the LC4 glomerulus (outlined with a green dashed line). Topographic separation of the LC4 axon terminals occurs along the sagittal diameter of the glomerulus. Normalized value of the sagittal diameter was used to assess the relative placement of the postsynaptic dendrites (see Methods). Dotted straight lines indicate the positions of DN dendritic centroids along the sagittal diameter of the glomerulus. Position of the LC4 glomerulus centroid slightly deviates from 0.5 value due to the naturally curved shape of the glomerulus. **d**, Strategy for sparse labeling of LC4 neurons and their presynaptic sites. Labeling of cell membranes (myr::GFP) and presynaptic sites (Brp-smGdP-V5) is dependent upon heat-shock induced expression of FLP (See Methods). **e**, Confocal projection of a single LC4 neuron with presynaptic sites labeled and colocalized with dendrite of DNp02 (n = 18 individual LC4 neurons from different brains, corresponding to Fig. 6e–h. **f**, Confocal projection of a single LPLC2 neuron with labeled presynaptic sites colocalized with GF dendrite (n = 10 individual LPLC2 neurons from different brains, corresponding to Fig. 6k,l). Regions corresponding to LC4 and LPLC2 glomeruli are indicated with dashed yellow rectangles.

Extended Data Fig. 11 The role of axo-dendritic overlap in synaptic specificity with and without axon topography.

Relationship between the number of synapses (VPNs and their postsynaptic targets) and axo-dendritic overlap score obtained from “hemibrain” EM volume using “overlap_score” function of the *natverse* package. **a**–**b**, representative examples of correlation plots featuring VPNs with (**a**) and without (**b**) topographic mapping in optic glomeruli (each dot represents a single neuron within one VPN cell type) and their individual postsynaptic partners. Measurements were performed using four resolution thresholds (3000, 1000, 300 and 100nm) to progressively distinguish the specific role

of general axonal topography (retinotopic mapping of axonal projections) from microtopography resulting from local synaptic specificity. An R^2 value reflects the impact of spatial positioning of axons and dendrites on the resulting synaptic connectivity. Error bands, SEM. D, dorsal; M, medial; P, posterior; L, lateral. **c**, plots summarizing the impact of topography on synaptic connectivity (at different spatial resolutions) for four VPN cell types with and without axonal topography and their five different postsynaptic partners (all examples are taken form Fig. 4 and Extended Data Fig. 7). **d**, Comparison of axo-dendritic overlap score for top10 individual VPN neurons with and without axonal topography (LC4, top and LPLC2, bottom, respectively, n = 10+10 for each VPN cell type) making most and least number of synapses with their postsynaptic targets (DNp02 and GF, respectively). Two-tailed unpaired Welch's *t*-test. In all box plots, the solid line depicts the median; the upper and lower bounds of the box depict the third and first quantiles of the data spread, respectively. Whiskers indicate min and max values.

Source data

Extended Data Fig. 12 Functional assessment of synaptic gradients in LPLC2 using electrophysiology.

a, Single focal plane snapshots of GFP expression by three tested VPN cell types overlaid with schematic ATP puffing locations. Focal application of 5 mM ATP (0.2 s pulse) was performed at either dorsal or ventral position along the dendrites of each VPN in the lobula. Outlines of the lobula (Lo) and lobula plate (LP), dashed white lines. D, dorsal; L, lateral. **b**, Depolarization responses of GF upon activation of VPNs from 6 different flies in each genotype (data from one animal shown in Fig. 6m). Blue, dorsal averaged responses. Red, ventral averaged responses. Grey, individual trials. **c**, Summary of early (top) and late (bottom) GF peak responses obtained in the time window of 0.5s and 8s after the stimulation onset, respectively. Mean differences between absolute values of GF responses to dorsal and ventral stimulations of the corresponding VPN cell types are shown next to each plot. Early peak: 3.32 ± 2.25 mV for LPLC2, 0.007 ± 0.06 mV for LC4, −0.04 ± 0.13 mV for LC11. Late peaks: 10.27 ± 1.88 mV for LPLC2, −1.12 ± 1.16 mV for LC4, −0.23 ± 0.52 mV for LC11 (error bars, SEM; n = 7

animals for each genotype. Individual data points are means of n = 5 trials per animal; two-tailed paired t-test). N.S.: $P > 0.05$; *: $P < 0.05$; **: $P < 0.01$; ****: $P < 0.0001$. Detailed description of statistical tests and p-values for panels is available in Supplementary Table 1.

Source data

Supplementary information

Supplementary Information

This file contains Supplementary Tables 1 and 2 and legends for Supplementary Videos 1–6.

Reporting Summary

Supplementary Video 1

Supplementary Video 2

Supplementary Video 3

Supplementary Video 4

Supplementary Video 5

Supplementary Video 6

Source data

Source Data Figs. 1–4 and 6 and Extended Data Figs. 1–4, 9, 11 and 12

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Dombrovski, M., Peek, M.Y., Park, JY. *et al.* Synaptic gradients transform object location to action. *Nature* **613**, 534–542 (2023). https://doi.org/10.1038/s41586-022-05562-8

- Received: 13 May 2022
- Accepted: 11 November 2022
- Published: 04 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05562-8

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

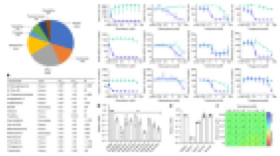
Provided by the Springer Nature SharedIt content-sharing initiative

A neural strategy for directional behaviour

- Daniel Tomsic
- Jamie Theobald

Nature News & Views 04 Jan 2023

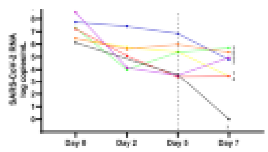
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05562-8



Pyrimidine inhibitors synergize with nucleoside analogues to block SARS-CoV-2

07 February 2022

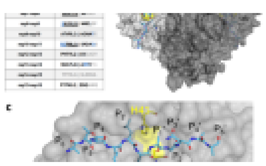
David C. Schultz, Robert M. Johnson, … Sara Cherry



A proof-of-concept study on the genomic evolution of Sars-Cov-2 in molnupiravir-treated, paxlovid-treated and drug-naïve patients

15 December 2022

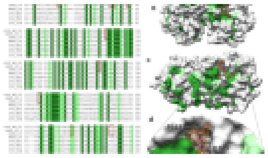
Claudia Alteri, Valeria Fox, … Carlo Federico Perno



Defining the substrate envelope of SARS-CoV-2 main protease to predict and avoid drug resistance

21 June 2022

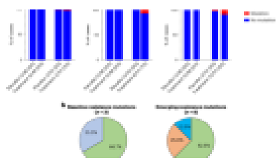
Ala M. Shaqra, Sarah N. Zvornicanin, … Celia A. Schiffer



[A small molecule compound with an indole moiety inhibits the main protease of SARS-CoV-2 and blocks virus replication]

28 January 2021

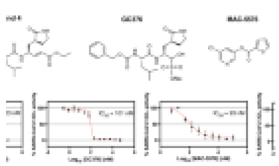
Shin-ichiro Hattori, Nobuyo Higashi-Kuwata, … Hiroaki Mitsuya



[Emergence of SARS-CoV-2 escape mutations during Bamlanivimab therapy in a phase II randomized clinical trial]

26 October 2022

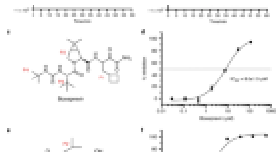
Manish C. Choudhary, Kara W. Chew, … for the ACTIV-2/A5401 Study Team



[Development of optimized drug-like small molecule inhibitors of the SARS-CoV-2 3CL protease for treatment of COVID-19]

07 April 2022

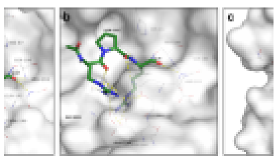
Hengrui Liu, Sho Iketani, … Brent R. Stockwell



[Both Boceprevir and GC376 efficaciously inhibit SARS-CoV-2 by targeting its main protease]

04 September 2020

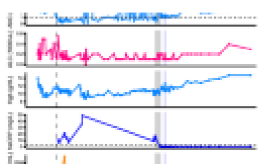
Lifeng Fu, Fei Ye, … George Fu Gao



[Peptidomimetic inhibitors of TMPRSS2 block SARS-CoV-2 infection in cell culture]

08 July 2022

Lukas Wettstein, Philip Maximilian Knaff, … Volker Mailänder



[De novo emergence of a remdesivir resistance mutation during treatment of persistent SARS-CoV-2 infection in an immunocompromised patient: a case report]

17 March 2022

Shiv Gandhi, Jonathan Klein, … Albert I. Ko

- Article
- Open Access

- Published: 09 November 2022

Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir

- Sho Iketani,
- Hiroshi Mohri,
- Bruce Culbertson,
- Seo Jung Hong,
- Yinkai Duan,
- Maria I. Luck,
- Medini K. Annavajhala,
- Yicheng Guo,
- Zizhang Sheng,
- Anne-Catrin Uhlemann,
- Stephen P. Goff,
- Yosef Sabo,
- Haitao Yang,
- Alejandro Chavez &
- …
- David D. Ho

Show authors

Nature **volume 613**, pages 558–564 (2023)

- 14k Accesses
- 6 Citations
- 35 Altmetric
- Metrics details

Abstract

Nirmatrelvir, an oral antiviral targeting the 3CL protease of SARS-CoV-2, has been demonstrated to be clinically useful against COVID-19 (refs. [1,2]). However, because SARS-CoV-2 has evolved to become resistant to other therapeutic modalities[3,4,5,6,7,8,9], there is a concern that the same could occur for nirmatrelvir. Here we examined this possibility by in vitro passaging of SARS-CoV-2 in nirmatrelvir using two independent approaches, including one on a large scale. Indeed, highly resistant viruses emerged from both and their sequences showed a multitude of 3CL protease mutations. In the experiment peformed with many replicates, 53 independent viral lineages were selected with mutations observed at 23 different residues of the enzyme. Nevertheless, several common mutational pathways to nirmatrelvir resistance were preferred, with a majority of the viruses descending from T21I, P252L or T304I as precursor mutations. Construction and analysis of 13 recombinant SARS-CoV-2 clones showed that these mutations mediated only low-level resistance, whereas greater resistance required accumulation of additional mutations. E166V mutation conferred the strongest resistance (around 100-fold), but this mutation resulted in a loss of viral replicative fitness that was restored by compensatory changes such as L50F and T21I. Our findings indicate that SARS-CoV-2 resistance to nirmatrelvir does readily arise via multiple pathways in vitro, and the specific mutations observed herein form a strong foundation from which to study the mechanism of resistance in detail and to inform the design of next-generation protease inhibitors.

Main

The COVID-19 (coronavirus disease, 2019) pandemic has continued to affect the global populace. The rapid development and deployment of effective vaccines, as well as monoclonal antibody therapeutics beginning in late 2020, have helped greatly to curtail its impacts[10,11,12,13,14,15,16]. Nevertheless, the aetiologic agent, SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), has continuously evolved to develop resistance to antibody-mediated neutralization[4,5,6,7,8]. In particular, several recent Omicron subvariants exhibit such strong antibody resistance that vaccines have had their protection against infection dampened and a majority of current monoclonal therapeutics have lost efficacy[4,5,8], as manifested by

increasing levels of breakthrough infections in convalescing and/or vaccinated individuals[3].

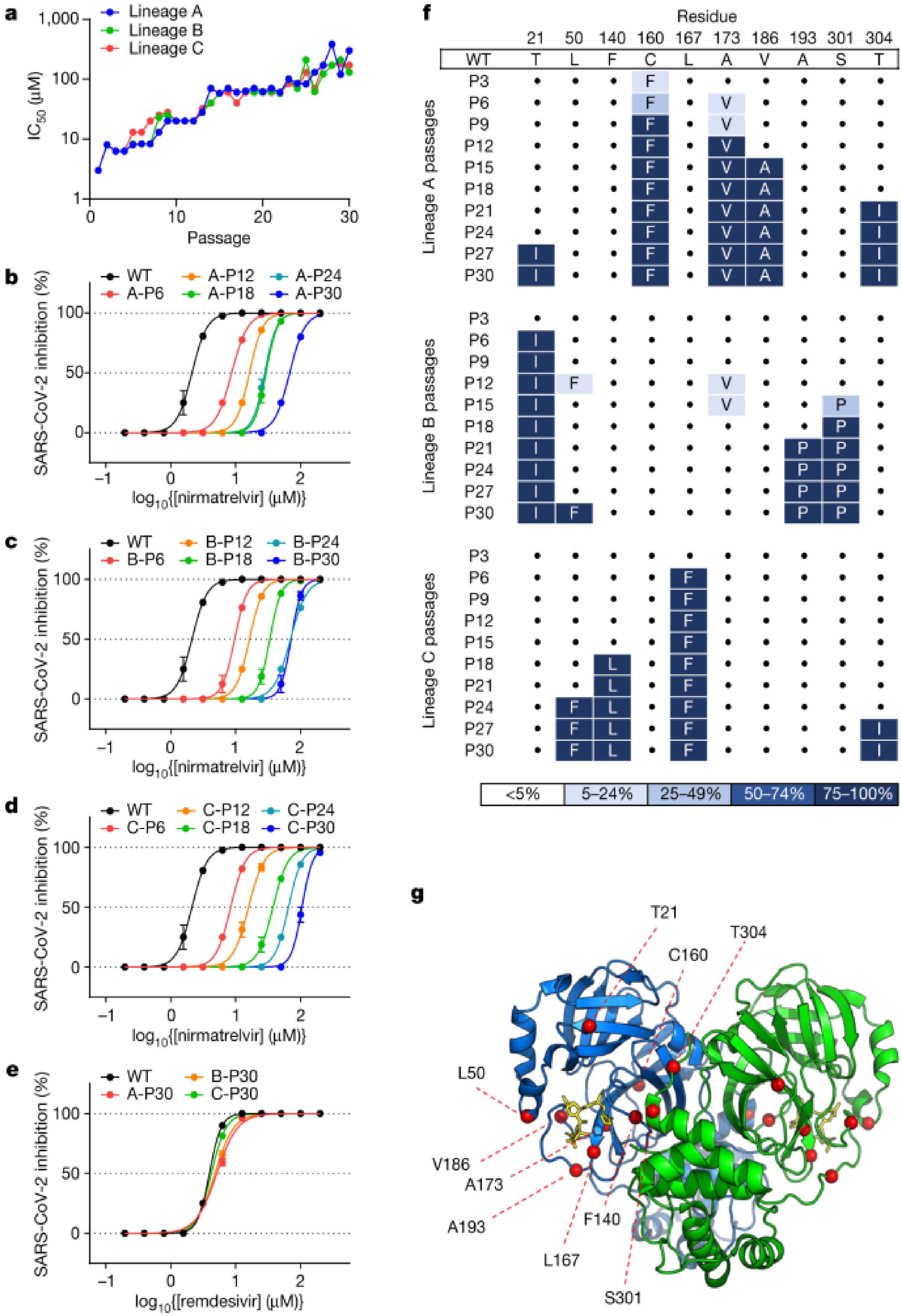
Fortunately, treatment options remain. In the United States, three antivirals have received emergency use authorization for COVID-19 treatment: remdesivir[17,18], molnupiravir[19,20,21] and nirmatrelvir[1,2] (also known as PF-07321332, used in combination with ritonavir and marketed as PAXLOVID). The first two target the RNA-dependent RNA polymerase (RdRp) and the latter targets the 3CL protease ($3CL^{pro}$, also known as main protease (M^{pro}) and nonstructural protein 5 (nsp5)). Both enzymes are essential for the viral life cycle and are relatively conserved among coronaviruses[22,23]. Remdesivir is administered intravenously and has a reported relative risk reduction of 87% (ref. [18]), whereas molnupiravir and nirmatrelvir are administered orally and have reported clinical efficacies of 31% (ref. [20]) and 89% (ref. [1]), respectively, in lowering rates of hospitalization or death. As the use of these antivirals increases there is a concern that drug resistance may arise, particularly if given as monotherapies. For remdesivir, in vitro and in vivo studies have shown mutations associated with resistance[9,24,25], and resistance to molnupiravir or nirmatrelvir is now under active investigation. Here we report that there are multiple routes by which SARS-CoV-2 can gain resistance to nirmatrelvir in vitro.

Nirmatrelvir resistance in Vero E6

To select for resistance to nirmatrelvir, SARS-CoV-2 (USA-WA1/2020 strain) was passaged in the presence of increasing concentrations of the drug (Methods). We conducted this initial experiment in triplicate, using Vero E6 cells because they have been one of the standard cell lines used in COVID-19 research. After 30 passages each of the three lineages demonstrated a high level of resistance, with half-maximal inhibitory concentration (IC_{50}) values increasing 33- to 50-fold relative to that of the original virus (Fig. 1a–d). Examination of earlier viral passages confirmed a stepwise increase in nirmatrelvir resistance with successive passaging (Fig. 1b–d), with no evidence of resistance to remdesivir (Fig. 1e). The resistant viruses selected by passaging maintained their replicative fitness in vitro, with growth

kinetics similar to those passaged without nirmatrelvir (Extended Data Fig. 1).

Fig. 1: Identification of nirmatrelvir resistance in Vero E6 cells.


a
Lineage A
Lineage B
Lineage C
IC50 (μM)
1,000
100
10
1
Passage
0
10
20
30
b
WT
A-P6
A-P12
A-P18
A-P24
A-P30
SARS-CoV-2 inhibition (%)
log10{[nirmatrelvir] (μM)}
c
WT
B-P6
B-P12
B-P18
B-P24
B-P30
SARS-CoV-2 inhibition (%)
log10{[nirmatrelvir] (μM)}
d
WT
C-P6
C-P12
C-P18
C-P24
C-P30
SARS-CoV-2 inhibition (%)
log10{[nirmatrelvir] (μM)}
e
WT
A-P30
B-P30
C-P30
SARS-CoV-2 inhibition (%)
log10{[remdesivir] (μM)}
f
Residue
21 50 140 160 167 173 186 193 301 304
WT T L F C L A V A S T
Lineage A passages
Lineage B passages
Lineage C passages
P3 P6 P9 P12 P15 P18 P21 P24 P27 P30
<5%
5–24%
25–49%
50–74%
75–100%
g
T21
T304
C160
L50
V186
A173
A193
F140
L167
S301

a, Changes in IC_{50} during passaging of SARS-CoV-2 with nirmatrelvir. Vero E6 cells were infected in triplicate with SARS-CoV-2 (USA-WA1/2020) and passaged to fresh cells every 3 days for 30 passages (Methods). **b**–**d**, Validation of nirmatrelvir resistance for the indicated passage from each of the three lineages (A (**b**), B (**c**) and C (**d**)). **e**, Inhibition of passage 30 viruses from each lineage by remdesivir. **f**, Mutations in $3CL^{pro}$ found in the indicated passages from each lineage. Dots indicate WT at that residue. Mutations are shaded according to frequency. **g**, Residues mutated with passaging in Vero E6 cells overlaid onto the $3CL^{pro}$ structure with nirmatrelvir bound. The Cα of each mutated residue is denoted by a red sphere. The $3CL^{pro}$–nirmatrelvir complex was downloaded from PDB under accession code 7VH8. **a**–**e**, Error bars denote mean ± s.e.m of four technical replicates.

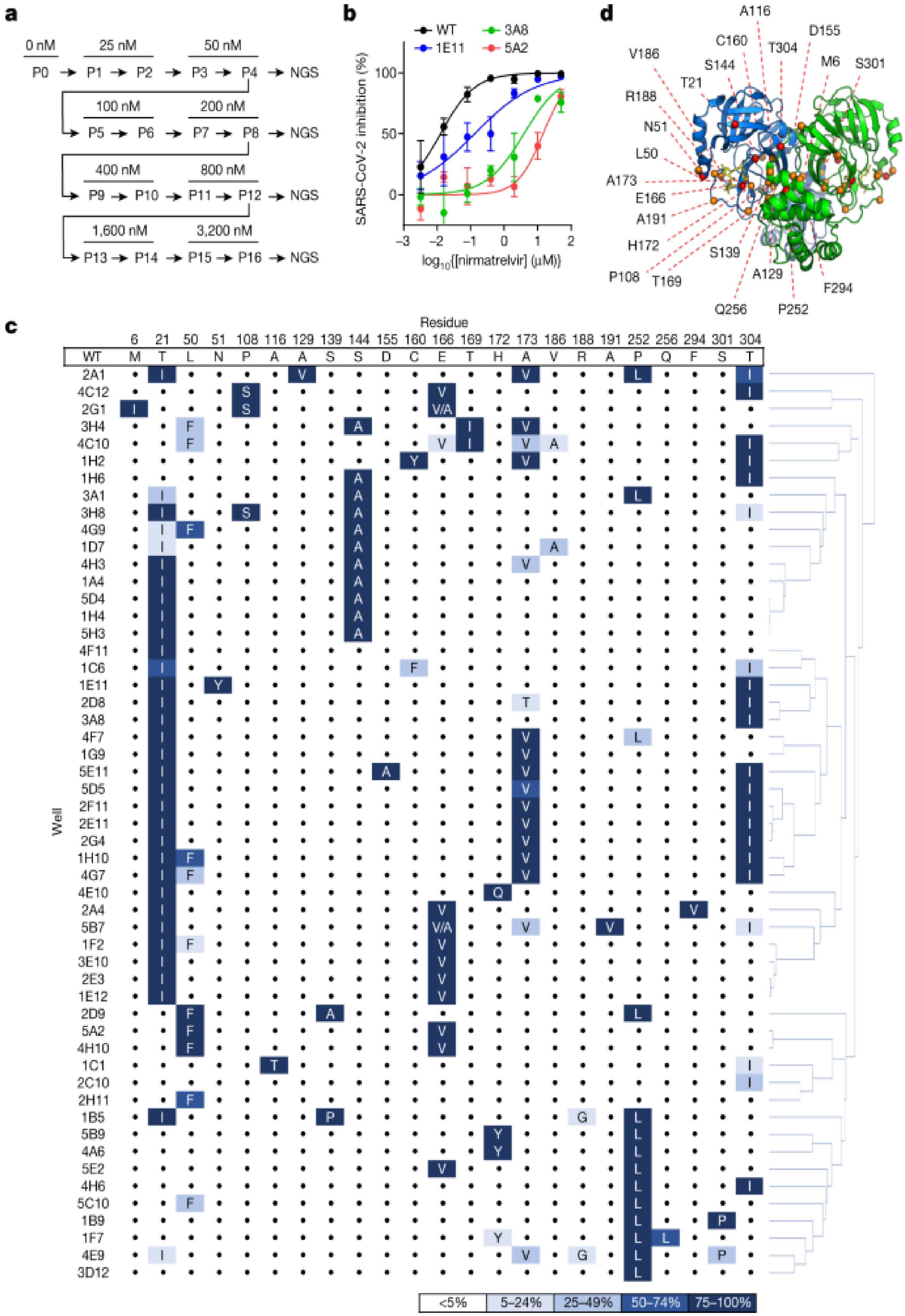
We then sequenced the $3CL^{pro}$ gene from the three viral lineages collected every three passages to investigate which mutations might confer resistance (Fig. 1f). We found that the three lineages harboured unique mutations with only one mutation, at most, overlapping between the different lineages (T21I in lineages A and B, L50F in lineages B and C and T304I in lineages A and C). The observed mutations occurred in a stepwise manner, mirroring the increases in drug resistance (Fig. 1f) and a number of them, but not all, were situated near the nirmatrelvir-binding site (Fig. 1g). Specifically, F140L and L167F were within 5 Å of nirmatrelvir. These results suggested that SARS-CoV-2 could readily develop nirmatrelvir resistance using any one of several mutational pathways.

Nirmatrelvir resistance in Huh7-ACE2 cells

We therefore set out to conduct another passaging experiment to select for nirmatrelvir resistance, but this time at a larger scale with many replicates to better capture the multitude of solutions that SARS-CoV-2 could adopt under drug pressure. For these later studies, we utilized Huh7-ACE2 cells to examine whether differences would arise in human cells, and because Vero E6 cells express high levels of P-glycoprotein, an efflux transporter that limits the intracellular accumulation of nirmatrelvir[26]. We passaged SARS-CoV-2-mNeonGreen (USA-WA1/2020 background with ORF7 replaced by

mNeonGreen[27]) independently in 480 wells for 16 passages with increasing concentrations of nirmatrelvir over time, and viruses from every fourth passage were subjected to next-generation sequencing (NGS) (Fig. 2a and Methods). After 16 passages, varying degrees of nirmatrelvir resistance were observed as exemplified by the three viruses shown in Fig. 2b. Sequencing of 3CL^{pro} in all wells that retained mNeonGreen signal identified 53 mutant populations (Fig. 2c). Across all of these populations, mutations were observed at 23 residues within the enzyme (between one and six mutations in each isolate), both proximal (at least 5 Å; S144A, E166(A/V), H172(Q/Y) and R188G) and distal (over 5 Å) to nirmatrelvir (Fig. 2d). Whereas there was widespread diversity among the passaged populations, seven mutations appeared ten or more times across replicates: T21I, L50F, S144A, E166V, A173V, P252L and T304I. The only 3CL^{pro} cleavage site mutation frequently observed was T304I, which corresponds to the cleavage site nsp5/6 T(P3)I. Other sites were only rarely observed to mutate, suggesting that substrate cleavage site alterations are largely not responsible for nirmatrelvir resistance (Extended Data Fig. 2), with the possible exception of *cis*-cleavage.

Fig. 2: Identification of nirmatrelvir resistance at scale in Huh7-ACE2 cells.

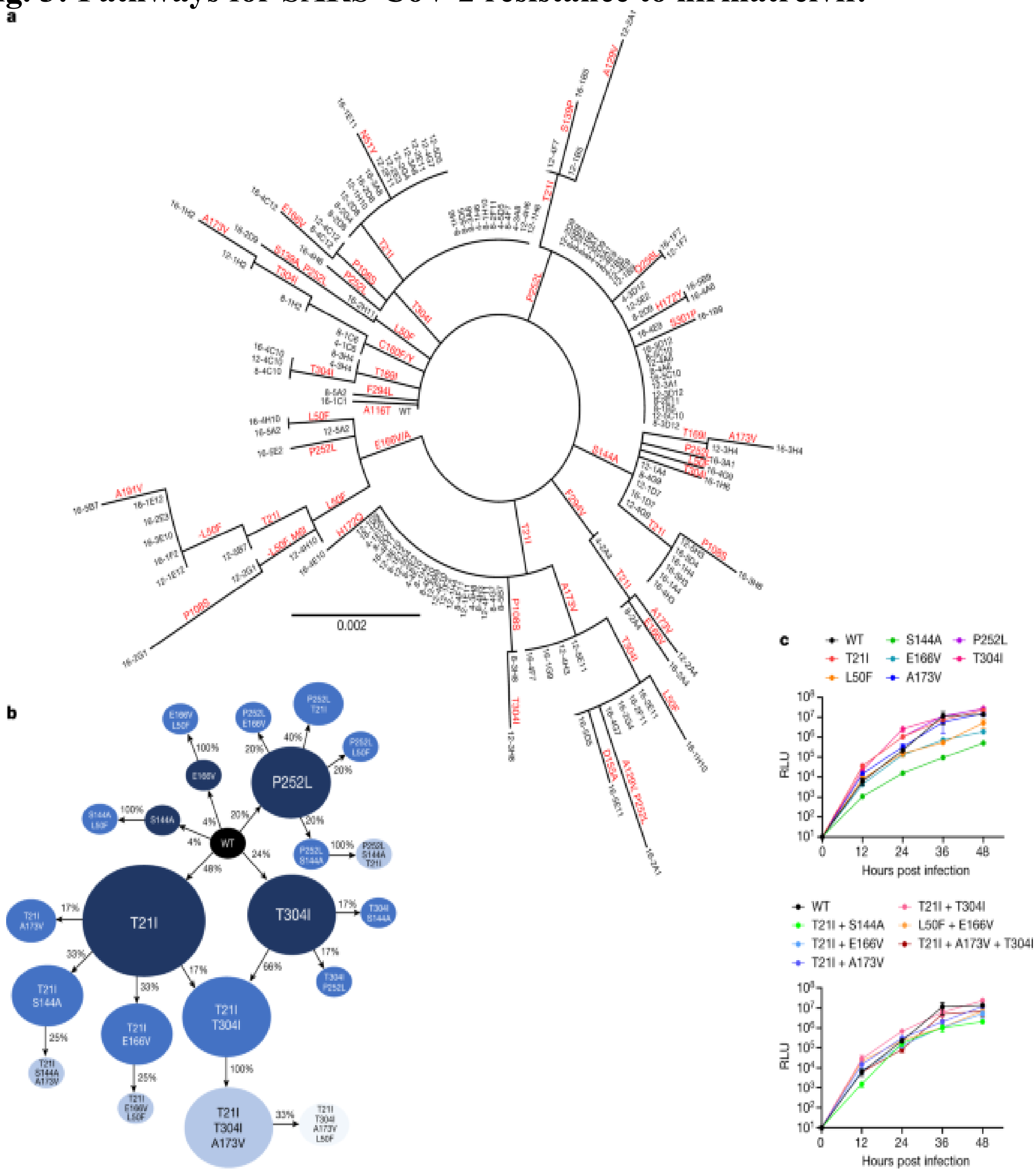


a
0 nM
25 nM
50 nM
P0 → P1 → P2 → P3 → P4 → NGS
100 nM
200 nM
P5 → P6 → P7 → P8 → NGS
400 nM
800 nM
P9 → P10 → P11 → P12 → NGS
1,600 nM
3,200 nM
P13 → P14 → P15 → P16 → NGS
b
WT
1E11
3A8
5A2
SARS-CoV-2 inhibition (%)
100
50
0
−3
−2
−1
0
1
2
log10([nirmatrelvir] (μM))
d
A116
D155
C160
T304
V186
S144
M6
S301
T21
R188
N51
L50
A173
E166
A191
H172
S139
P108
T169
A129
F294
Q256
P252
c
Residue
6 21 50 51 108 116 129 139 144 155 160 166 169 172 173 186 188 191 252 256 294 301 304
WT M T L N P A A S S D C E T H A V R A P Q F S T
Well
2A1 I V V L I
4C12 S V I
2G1 I S V/A
3H4 F A I V
4C10 F V I V A I
1H2 Y V I
1H6 A I
3A1 I A L
3H8 I S A I
4G9 I F A
1D7 I A A
4H3 I A V
1A4 I A
5D4 I A
1H4 I A
5H3 I A
4F11 I
1C6 I F I
1E11 I Y I
2D8 I T I
3A8 I I
4F7 I V L
1G9 I V
5E11 I A V I
5D5 I V I
2F11 I V I
2E11 I V I
2G4 I V I
1H10 I F V I
4G7 I F V I
4E10 I Q
2A4 I V V
5B7 I V/A V V I
1F2 I F V
3E10 I V
2E3 I V
1E12 I V
2D9 F A L
5A2 F V
4H10 F V
1C1 T I
2C10 I
2H11 F
1B5 I P G L
5B9 Y L
4A6 Y L
5E2 V L
4H6 L I
5C10 F L
1B9 L P
1F7 Y L L
4E9 I V G L P
3D12 L
<5%
5–24%
25–49%
50–74%
75–100%

a, Passaging scheme: 480 wells were infected with SARS-CoV-2-mNeonGreen and passaged to fresh Huh7-ACE2 cells every 3–4 days, with the concentration of drug doubled every two passages. **b**, Validation of nirmatrelvir resistance of three wells from passage 16. These viral populations had the following mutations: 3A8 (T21I, T304I), 1E11 (T21I, N51Y, T304I) and 5A2 (L50F, E166V). See Supplementary Table 1 for exact frequencies. Representative curves from a single experiment from two biologically independent experiments are shown. Error bars denote mean ± s.e.m of three technical replicates. **c**, Mutations in $3CL^{pro}$ found in passage 16 from 53 wells. Dots indicate WT at that residue. Mutations are shaded according to frequency. **d**, Residues mutated in passaging in Huh7-ACE2 cells overlaid onto the $3CL^{pro}$ structure with nirmatrelvir bound. All 23 mutated residues across all resistant populations are indicated for any individual isolate having between one and six mutations. The Cα of each residue that was mutated is denoted by a red sphere for mutations observed more than ten times, and is denoted by an orange sphere for mutations observed fewer than ten times. The $3CL^{pro}$–nirmatrelvir complex was downloaded from PDB under accession code 7VH8.

Sequencing of the same wells at earlier passages showed less diversity in $3CL^{pro}$, with totals of 11, 16 and 22 unique mutations detected across all populations from passages 4, 8 and 12, respectively (Supplementary Table 1). Because a standard phylogenetic analysis showed a rather complex stepwise order of acquisition of mutations for each passaged lineage (Fig. 3a), we more carefully analysed the order in which mutations arose across the various lineages (Methods and Supplementary Table 1) and generated a pathway network delineating the most common routes taken by SARS-CoV-2 in vitro to develop nirmatrelvir resistance (Fig. 3b and Supplementary Table 2). The majority of these viral lineages descended initially from T21I, P252L and T304I, suggesting that these mutations may serve as ‘founder’ or ‘precursor’ mutations when drug concentrations are relatively low. Additional mutations then occurred, probably to increase the level of resistance as drug concentrations were increased and/or to compensate for reduced viral fitness. These findings indicated that, although there are multiple means by which SARS-CoV-2 can resist nirmatrelvir, several common mutational pathways are favoured.

Fig. 3: Pathways for SARS-CoV-2 resistance to nirmatrelvir.



a, Phylogenetic tree of sequences from passaging in Huh7-ACE2 cells. Only sequences with mutations are shown. Sequences are denoted as passage number, followed by the well number. Mutations that arose along particular branches are annotated in red; ‘-’ denotes when a mutation appears to have been lost from a particular branch. **b**, Observed pathways for nirmatrelvir resistance in Huh7-ACE2 cells. The most commonly observed mutations in passage 16 were used to build these pathways (Methods and Supplementary

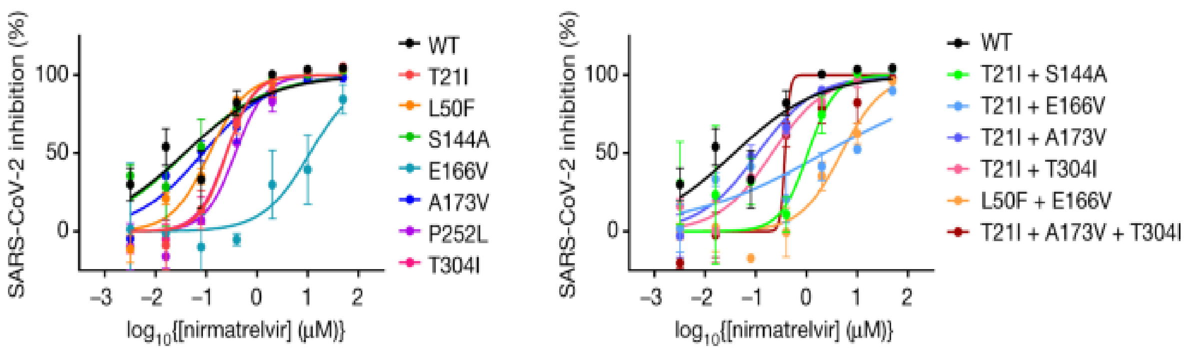
Table 2). Nodes are shaded from dark to light, with founder mutations darker. Percentages indicate the frequency by which child nodes derive from the immediate parental node. Descendent arrows that do not sum to 100% indicate that a proportion did not advance beyond the indicated mutations in the experiment. **c**, Growth assay with recombinant live SARS-CoV-2 carrying single (top) and combination 3CL^{pro} mutations (bottom). Huh7-ACE2 cells were infected with 0.01 multiplicity of infection (MOI) of virus, and luminescence was quantified at the indicated time points. S144A, E166V and T21I + S144A are statistically significant from WT at 48 h (two-way analysis of variance with Geisser–Greenhouse correction followed by Dunnett's multiple comparisons test; $P = 0.0039$, $P = 0.0006$, $P = 0.0006$, respectively). Representative curves from a single experiment from two biologically independent experiments are shown. Error bars denote mean ± s.e.m of three technical replicates. RLU, relative luminescence units.

Nirmatrelvir resistance mutations

To further investigate which mutations were responsible for nirmatrelvir resistance we proceeded to generate recombinant SARS-CoV-2 clones, each containing a unique mutation or a combination of mutations. To construct the 15 mutant viruses from the first passage experiment (Fig. 1f) and the 22 mutant viruses from the second (Fig. 3a,b) would be beyond the scope of the current study. We therefore decided to focus on the seven most common single-point mutants from the large passaging study, as well as on five double mutants and one triple mutant (Extended Data Fig. 3). All viruses grew similarly to wild type (WT) in the absence of drug except for S144A, E166V and T21I + S144A, which were significantly impaired in their growth kinetics (Fig. 3c). However, both T21I + E166V and L50F + E166V replicated well with kinetics similar to WT, suggesting that T21I and L50F each compensated for the fitness loss of E166V. Of the individual mutants tested against nirmatrelvir, E166V was most resistant (100-fold) with P252L and T304I having low-level resistance (around sixfold) and S144A and A173V only minimal resistance (about threefold or less) (Fig. 4a,b and Extended Data Fig. 4). Combination of either T21I or L50F with E166V resulted in a virus that was substantially resistant to nirmatrelvir (83- and 53-fold, respectively), but with WT replicative kinetics (Fig. 3c).

Fig. 4: Validation of identified mutations in isogenic recombinant SARS-CoV-2.

a



b

Fold change in IC_{50} relative to WT	Nirmatrelvir	Ensitrelvir	Remdesivir
T21I	−4.6	−1.7	1.7
L50F	−4.2	−2.8	1.2
S144A	−2.2	−13	2.9
E166V	−100	−23	3.7
A173V	−1.7	−1.7	−2.0
P252L	−5.9	−1.9	1.5
T304I	−5.5	−1.6	1.6
T21I + S144A	−9.4	−18	1.7
T21I + E166V	−83	−3.4	−2.4
T21I + A173V	−3.1	2.0	−2.1
T21I + T304I	−3.3	1.0	2.8
L50F + E166V	−53	−3.7	−1.9
T21I + A173V + T304I	−15	2.1	−2.9

<−5	<−10	<−100

a, Individual inhibition curves of recombinant live SARS-CoV-2 carrying single (left) and combination $3CL^{pro}$ mutations (right) by nirmatrelvir. Representative curves from a single experiment from three biologically independent experiments are shown. Error bars denote mean ± s.e.m of three technical replicates. **b**, Inhibition of recombinant live SARS-CoV-2 carrying single and combination $3CL^{pro}$ mutations by nirmatrelvir, ensitrelvir and

remdesivir. Values shown are fold change of mean values in IC_{50} relative to inhibition of WT from three biologically independent experiments.

We next tested this panel of viruses against ensitrelvir[28] (also known as S-217622), another 3CL protease inhibitor that has demonstrated clinical efficacy[29], for cross-resistance, together with remdesivir as control. Only S144A, E166V and T21I + S144A showed substantial (13- to 23-fold) cross-resistance to ensitrelvir (Fig. 4b and Extended Data Figs. 4 and 5). As expected, none of these mutations conferred resistance to remdesivir. We additionally tested the passage 30 viruses resulting from the initial selection experiment in Vero E6 cells (Fig. 1) against these two inhibitors. Again, all three lineages were as susceptible to remdesivir as WT and only lineage C (L50F + F140L + L167F + T304I) showed cross-resistance to ensitrelvir (approximately 25-fold) (Extended Data Fig. 6). This may be due to F140L, because L50F and T304I did not demonstrate ensitrelvir resistance (Fig. 4b) and L167 does not contact ensitrelvir (see below). Together these results suggest that some mutations, such as E166V, can confer a high degree of nirmatrelvir resistance alone whereas others, such as T21I, P252L and T304I, confer only low levels of resistance individually. The degree of cross-resistance to ensitrelvir was variable among the tested mutant viruses, probably due to the differences in binding of these drugs to the substrate-binding site of $3CL^{pro}$ (Extended Data Fig. 7a). Nevertheless, it is clear that selection for nirmatrelvir resistance can yield mutations that confer cross-resistance to other inhibitors of clinical interest as well.

To begin to understand the mechanisms underlying the resistance conferred by these mutations, we considered their structural context. Nirmatrelvir and ensitrelvir both bind within the substrate-binding site, but in differing modes, which may have resulted in the differences observed in the inhibition profiles of the mutants (Fig. 4b and Extended Data Fig. 7a). E166 directly interacts with the lactam ring of nirmatrelvir via hydrogen bonding, and the valine substitution at this position may abrogate some of these interactions to result in the strong drug resistance observed (Extended Data Fig. 7b). E166 is also able to form hydrogen bonds with the first residue (S1) of the neighbouring protomer and is therefore involved in dimerization, which is essential for protease activity because $3CL^{pro}$ functions as a homodimer[30]. The disruption of hydrogen-bonding interactions (Extended Data Fig. 7b)

may explain the reduced fitness of the E166V mutant (Fig. 3c). The side chain of S144 forms a hydrogen bond with the main chain of L141 to stabilize the S1 subsite of the substrate-binding site, so the S144A mutation may disorder this region and hamper the binding of both nirmatrelvir and ensitrelvir (Extended Data Fig. 7c), although it is not clear why this requires the T21I mutation in conjunction. L167 participates in the formation of the S4 subsite, and the L167F mutation may cause a steric clash with nirmatrelvir (Extended Data Fig. 7d). However, because ensitrelvir does not extend into the S4 subsite, this mutation may not be responsible for the cross-resistance observed in lineage C (Extended Data Fig. 6). Because F140 interacts by π-π stacking interactions with H163, which directly interacts with both nirmatrelvir and ensitrelvir, the F140L mutation may abrogate this interaction, resulting in resistance (Extended Data Figs. 6 and 7a,b). For a number of these mutations, however, it is not immediately apparent how they confer drug resistance given that they are distant from the substrate-binding site where the drugs bind (Extended Data Fig. 3).

Finally, we compared the mutations identified in this study with clinical SARS-CoV-2 sequences reported to the Global Initiative on Sharing Avian Influenza Data[31]. Nearly all of the mutations we have identified were observed among viruses circulating in the population, albeit at low frequencies (Extended Data Fig. 8a). Comparing the frequencies of these mutations in periods before and after authorization of the combination of nirmatrelvir and ritonavir (PAXLOVID) did not show an appreciable increase in the observed mutations (Extended Data Fig. 8b).

Discussion

Because antibody-based interventions for SARS-CoV-2 face increasing resistance from the emergence of variants of concern, antivirals with alternative modes of action have increased in importance. Nirmatrelvir, as an oral antiviral targeting $3CL^{pro}$, is a therapeutic that has shown high efficacy in lowering severe disease and hospitalization in infected persons who are at high risk and not vaccinated[1,2]. Indeed, it is the antiviral drug most commonly used to treat COVID-19 today[32]. Given the adaptations that the virus has already exhibited to other modes of treatment[3,4,5,6,7,8,9], it is clinically important to understand the mechanisms by which nirmatrelvir

resistance can occur. The results presented herein demonstrate that in vitro high-level resistance to nirmatrelvir can readily be achieved by SARS-CoV-2, and that this can occur in a multitude of ways. This finding is consistent with our previous report on the extensive plasticity of $3CL^{pro}$, as discovered by deep mutational scanning[33].

In both Vero E6 cells (Fig. 1) and Huh7-ACE2 cells (Fig. 2), multiple lineages with nonoverlapping mutations evolved under increasing drug pressure, consistent with that seen in similar small-scale studies[24,25,34,35]. Conducting selection at scale, however, showed that there are multiple mutational pathways to nirmatrelvir resistance but with several common trajectories preferred (Figs. 2c and 3a,b). A majority of lineages descended from viruses that acquired T21I, P252L or T304I as an initial mutation. Recombinant SARS-CoV-2, constructed to contain each of these point mutants, exhibited low-level resistance (Fig. 4a,b), suggesting that each of these precursor mutations may have allowed the virus to tolerate low concentrations of nirmatrelvir but required additional mutations as drug pressure was increased. Notably, all three of these mutations are somewhat distal (over 5 Å) from nirmatrelvir (Fig. 2d) and their mechanism for resistance is not evident without additional studies. We note, however, that T304 corresponds to the P3 site on the nsp5/6 cleavage substrate for $3CL^{pro}$ of both SARS-CoV and SARS-CoV-2 (Extended Data Fig. 2). Although the P3 site is exposed to solvent and is thus not considered to confer stringent substrate specificity, it has been shown that a suitable functional group (such as the side chain of isoleucine) at the P3 site can assist in increasing inhibitor/substrate potency and selectivity for $3CL^{pro}$ (refs. [36,37,38]). Therefore, it is possible that T304I could facilitate binding of the nsp5/6 cleavage site or promote the autocleavage process. The differing mutations observed between the two cell lines further emphasize the complexity and variety of pathways leading to nirmatrelvir resistance, although it is not yet clear why certain mutations are specific to the Vero E6 cell line.

Analyses with isogenic mutants also showed that several mutations are responsible for the observed nirmatrelvir resistance, with the E166V mutation conferring the most resistance (100-fold) (Fig. 4b), as reported elsewhere[33,35]. This mutation also conferred a degree of cross-resistance to ensitrelvir, another clinically relevant $3CL^{pro}$ inhibitor[28,29]. The mechanism

of resistance of E166V can be explained because it resides in the substrate-binding site, and valine substitution disrupts its hydrogen bonding to the lactam ring of nirmatrelvir (Extended Data Fig. 7b). However, this mutation lowered the replicative fitness of the virus in vitro (Fig. 3c), perhaps because of a loss of interaction with the first residue of the neighbouring protomer in dimerization (Extended Data Fig. 7b)[30]. Importantly, replicative fitness was restored when T21I or L50F was added (Fig. 3c), with no significant impact on drug resistance (Fig. 4b). How these two mutations compensate for the fitness loss of E166V remains unknown. It is worth mentioning that the E166V mutation was reportedly found in viral isolates from several PAXLOVID-treated individuals in the EPIC-HR clinical trial[1] (see Fact Sheet for Healthcare Providers: Emergency Use Authorization for PAXLOVID, revised 6 July 2022 (ref. [39])).

We have also found that a number of additional mutations could confer resistance to nirmatrelvir in vitro. T21I + S144A mediated not only significant resistance to nirmatrelvir but also cross-resistance to ensitrelvir (Fig. 4b), but this virus exhibited slower growth kinetics (Fig. 3c). Likewise, we inferred that both L167F and F140L were probably mediating drug resistance in the C-P30 lineage of the first in vitro passaging experiment (Fig. 1f) as discussed above, along with possible structural explanations. It is clear, nevertheless, that we have studied only a limited number of the mutational pathways taken by SARS-CoV-2 to evade nirmatrelvir. Furthermore, many of the mutations shown by our study are without a straightforward structural explanation at this time, and indeed, whereas other in vitro or in silico studies have identified residues such as E166 to be of importance, they have missed these other residues that are distant from the substrate-binding site[40,41,42]. It should also be mentioned that our studies were conducted with the ancestral WA1 strain, and the currently circulating Omicron variants—all of which except for BA.3 contain a P132H mutation in $3CL^{pro}$—may differ in their nirmatrelvir evasion pathways. Whereas this mutation has been reported to have no direct effect on nirmatrelvir resistance, it may influence the emergence of subsequent resistance-conferring mutations[43]. It will require extensive virological, biochemical and structural studies to delineate which mutations confer resistance and how, as well as to understand how certain mutations play compensatory roles. A better understanding of the mechanisms of nirmatrelvir resistance could

provide insight into the development of the next generation of $3CL^{pro}$ inhibitors.

At the time of writing, nirmatrelvir has been used to treat COVID-19 for only 6 months or less in most countries. SARS-CoV-2 resistance to this drug in patients has yet to be reported, and we see no appreciable difference in frequencies of the $3CL^{pro}$ mutations that we have uncovered in periods before and after emergency use authorization (Extended Data Fig. 8). Perhaps the absence of nirmatrelvir resistance in patients to date is due to the high drug concentrations achieved with the prescribed regimen, making it difficult for the virus to accumulate mutations in a stepwise manner. In addition, the drug is administered while the immune system is also actively eliminating the virus, including any resistant forms that may have emerged. Therefore, it makes sense to focus our surveillance effort on immunocompromised individuals on nirmatrelvir treatment for the appearance of drug-resistant virus. Past experience with other viral infections tells us that if drug resistance can be selected in vitro, it surely will occur also in vivo. Although current COVID-19 therapies have been largely administered as monotherapies, it is possible that future treatment will benefit from the use of a combination of drugs to minimize the likelihood of SARS-CoV-2 escape.

Methods

Biosafety

All SARS-CoV-2 passaging, infection and recombinant virus production was conducted in BSL-3 laboratories at Columbia University Irving Medical Center under procedures and guidelines approved by the Columbia University Institutional Biosafety Committee.

Compounds

Nirmatrelvir was purchased from Aobius, ensitrelvir from Glixx Laboratories and remdesivir from Selleckchem.

Cells

Vero E6 cells were obtained from ATCC (no. CRL-1586), HEK293T cells from ATCC (no. CRL-3216) and Vero E6-TMPRSS2-T2A-ACE2 cells from BEI Resources (no. NR-54970). Huh7-ACE2 cells were generated previously[33,44]. Cell morphology was visually confirmed before use and all cell lines tested mycoplasma negative. All cells were maintained at 37 °C under 5% CO_2.

In vitro selection for SARS-CoV-2 resistance to nirmatrelvir in Vero E6 cells

To select for the development of drug resistance against nirmatrelvir, WA1 (SARS-CoV-2, USA-WA1/2020 strain) was cultured in the presence of increasing concentrations of nirmatrelvir and passaged 30 times. Virus isolates recovered from culture at various passages were then characterized for their resistance to nirmatrelvir and their replication capacity.

To initiate passaging, Vero E6 cells were seeded in a 24-well plate at a density of 1×10^5 cells per well in complete medium (DMEM + 10% fetal calf serum + penicillin/streptomycin), and both drug and virus were then added the following day. The drug was prepared in a threefold dilution series based on its original IC_{50}. The virus was added at 5,000 50% tissue culture infectious dose ($TCID_{50}$) per well. Three days post infection, each well was scored for cytopathic effects (CPE) in a range of 0 to 4+ based on comparison with control wells as previously described[45], and 100 µl of the supernatant from the well with a CPE score equal to or greater than 2+ was passaged to each well in the next culture plate. The passage culture was set up in triplicate (lineages A, B and C) and passaging was performed independently—that is, viruses in lineage A were kept within the lineage A series of wells at every passage. Along with the cultures passaged with nirmatrelvir, WA1 was passaged without nirmatrelvir in two independent wells to serve as a passage control.

Values of IC_{50} for each lineage in passaging were determined based on CPE scores at day 3 of each passage; these values were derived using DeltaGraph

(Red Rock Software).

Sequencing of SARS-CoV-2 passaged in Vero E6 cells

For SARS-CoV-2 passaged in Vero E6 cells, passages were sequenced by either Sanger or Nanopore sequencing. For Sanger sequencing, viral RNA was isolated from culture supernatant with the QIAamp Viral RNA Mini Kit (Qiagen), reverse transcribed to complementary DNA with Superscript IV Reverse Transcriptase (ThermoFisher) and the priming primer (nsp5.R1) and subjected to nested PCR with Platinum SuperFi II (ThermoFisher) to obtain the full-length nsp5 gene. Primers for the first PCR were nsp5.F1: 5′-GTAGTGATGTGCTATTACCTCTTACGC-3′ and nsp5.R1: 5′-GCAAAAGCAGACATAGCAATAATACC-3′. Primers for the second PCR were nsp5.F2: 5′-CTTCAGTAACTCAGGTTCTGATGTTCT-3′ and nsp5.R2: 5′-ACCATTGAGTACTCTGGACTAAAACTAAA-3′. Both PCRs were run under the same conditions: 98 °C for 30 s, 25 cycles of 98 °C for 15 s, 60 °C for 10 s and 72 °C for 1 min, followed by 72 °C for 5 min. PCR products were purified and sequenced (Genewiz). Mixtures of viruses were determined by inspection of sequencing chromatograms. Sequences were analysed using Lasergene software (DNASTAR).

For Nanopore sequencing, viral RNA was isolated from culture supernatant with the QIAamp Viral RNA Mini Kit (Qiagen), Midnight RT PCR Expansion and Rapid Barcoding kits (Oxford Nanopore) were used for amplification and barcode-overlapping 1,200-base pair (bp) amplicons were tiled across the viral genome[46,47]. An Oxford Nanopore GridION with R9.4.1 flow cells was used for sequencing. Basecalling was performed in MinKNOW v.22.05.1. Consensus sequence generation was performed using the ONT Epi2Me ARTIC Nextflow pipeline v.0.3.16 (https://github.com/epi2me-labs/wf-artic). Pangolin 4.0.6 with UShER v.1.6 was used for parsimony-based lineage assignment. Sequences have been deposited at GenBank (nos. ON924329-ON924335 and ON930401-ON930431) (Supplementary Table 3).

Inhibition assay with SARS-CoV-2 passaged in Vero E6 cells

To characterize the inhibition of passaged viruses, each virus was first propagated in Vero E6 cells in the absence of drug and titrated using the Reed–Muench method[48]. Vero E6 cells were then seeded in 96-well plates at a density of 1.5×10^4 cells per well in complete medium. The following day, the virus was inoculated at a dose of 500 $TCID_{50}$ per well and a twofold dilution series of inhibitor added in quadruplicate. Three days post infection, the level of CPE was scored and IC_{50} was derived by fitting a nonlinear regression curve to the data in GraphPad Prism v.9.4 (Dotmatics).

Growth assay with SARS-CoV-2 passaged in Vero E6 cells

The fitness of passaged viruses was characterized by viral growth assay. Vero E6 cells were seeded in 96-well plates at a density of 1.5×10^4 cells per well in complete medium. The following day, the virus was inoculated at a dose of 200 $TCID_{50}$ per well in quadruplicate. At 6 h post infection, free virions in the culture were removed by changing of the medium twice. At 11, 24, 35 and 49 h post infection, 50 µl of culture supernatant from each well was collected and replenished with an equivalent volume of fresh medium. Viral RNA from each time point was purified using a PureLink Pro 96 Viral RNA/DNA Purification Kit (ThermoFisher), and viral copy number in each sample was then estimated by quantitative PCR with reverse transcription using TaqPath 1-Step RT–qPCR Master Mix (ThermoFisher) and a 2019-nCov CDC EUA Kit (Integrated DNA Technologies) with a 7500 Fast Dx Real-Time PCR Instrument (Applied Biosystems).

In vitro selection for SARS-CoV-2 resistance to nirmatrelvir in Huh7-ACE2 cells

To conduct selection at scale to observe as many resistance pathways as possible, SARS-CoV-2 infection was conducted in five 96-well plates thereby facilitating 480 independent selection lineages. We hypothesized that the use of limited number of cells would allow for a ‘bottleneck effect’ to occur, which would enable observation of rarer events that may be outcompeted from a larger population.

To initiate passaging, 3×10^4 Huh7-ACE2 cells per well were seeded in complete medium in five 96-well plates. The following day, all wells were infected with 0.05 MOI of SARS-CoV-2-mNeonGreen (a fluorescent reporter variant of USA-WA1/2020, gift of P.-Y. Shi)[27] without drug to generate passage 0. For each successive passage, cells were seeded the day before infection and the drug and virus then added 3–4 days after infection of the previous passage. The drug was initially added at 25 nM and then doubled every other successive passage. Viruses were transferred between passages by overlaying 50 µl of the supernatant from the previous passage. After 16 passages all 54 wells positive for mNeonGreen signal were sequenced, from which 53 lineages could be determined.

Inhibition assay with SARS-CoV-2 passaged in Huh7-ACE2 cells

To characterize the inhibition of passaged viruses, each of the viruses were first propagated in Huh7-ACE2 cells in the absence of drug and titrated using the Reed–Muench method[48]. Huh7-ACE2 cells were then seeded in 96-well plates at a density of 2×10^4 cells per well in complete medium. The following day, the virus was inoculated at a dose of 0.05 MOI per well and a fivefold dilution series of inhibitor added in triplicate. At 24 h post infection, supernatant was aspirated and cells were fixed with 4% paraformaldehyde in PBS and stained with DAPI. Cells were then imaged for DAPI and green fluorescent protein using IN Cell 2000 (GE) and analysed with CellProfiler v.4.0.7 (ref. [49]). The IC_{50} level was then derived by fitting a nonlinear regression curve to the data in GraphPad Prism v.9.4 (Dotmatics).

Sequencing of SARS-CoV-2 passaged in Huh7-ACE2 cells

For SARS-CoV-2 passaged in Huh7-ACE2 cells, passages were sequenced by Illumina NGS. Viral RNA was first extracted using a PureLink Pro 96 Viral RNA/DNA Purification Kit (ThermoFisher). Reverse transcription was carried out using a Maxima H Minus First Strand cDNA Synthesis Kit (ThermoFisher) with random hexamers according to the manufacturer's instructions. Briefly, 13.75 µl of viral RNA was mixed with 0.25 µl of random hexamers (50 ng μl^{-1}) and 1 µl of deoxynucleotide triphosphates

(dNTPs) (10 mM), and incubated at 65 °C for 5 min followed by 4 °C for 1 min. A mixture containing 4 µl of 5× RT buffer, 0.25 µl of enzyme mix (containing Maxima H Minus RT and RNase inhibitor) and 0.75 µl of H_2O was added to each sample and the reactions incubated at 25 °C for 10 min, 55 °C for 30 min and 85 °C for 5 min.

Sequencing libraries were prepared by amplification of either nine fragments tiled across the 3CL^{pro} open reading frame and adjacent nsp4/5 and nsp5/6 cut sites, or nine fragments containing each of the remaining 3CL^{pro} cut sites (see Supplementary Table 4 for primer sequences). Primers amplifying nonadjacent fragments of 3CL^{pro} were pooled and reactions carried out in technical duplicate, for a total of four first-round PCRs per sample. Each first-round PCR contained the following components: 1 µl of cDNA, 0.25 µl of 100 µM pooled primers, 0.4 µl of 10 mM dNTPs, 2 µl of 10× Taq buffer, 0.1 µl of Taq DNA polymerase (Enzymatics) and 16.25 µl of H_2O. Cycling conditions were as follows: (1) 94 °C for 3 min, (2) 94 °C for 30 s, (3) 54 °C for 20 s, (4) 72 °C for 30 s, (5) return to step 2 for 34 additional cycles, (6) 72 °C for 3 min and (7) hold at 4 °C.

Products from the four first-round PCRs for each sample were pooled and gel purified, and second-round indexing PCR was carried out for each sample with the following reagents: 1 µl of template DNA, 0.25 µl of each 100 µM indexing primer, 0.4 µl of 10 mM dNTPs, 2 µl of 10× Taq buffer, 0.1 µl of Taq DNA polymerase and 16.25 µl of H_2O. Cycling conditions were as follows: (1) 94 °C for 3 min, (2) 94 °C for 30 s, (3) 54 °C for 20 s, (4) 72 °C for 30 s, (5) return to step 2 for six additional cycles, (6) 72 °C for 3 min and (7) hold at 4 °C.

Second-round PCR products were pooled, gel purified and sequenced on an Illumina NextSeq system with 150-bp single-end reads. For select samples, sequences were confirmed using nanopore sequencing (Plasmidsaurus). For samples P16-2D9, P12-1A4 and 4-3A1, the original Illumina sequencing results were replaced by nanopore sequencing results.

For each sample, mutations and their frequencies were identified using the V-pipe computational pipeline (v.2.99.2)[50], with Wuhan-Hu-1 (GenBank accession no. MN908947) set as the reference sequence. Frequency

thresholds for reporting mutations were set at 5 and 10% for Illumina and nanopore sequencing, respectively. Supplementary Table 1 shows the absolute frequencies of mutations within each sample. Raw sequencing data have been deposited with NCBI Short Read Archive under BioProject Accession ID PRJNA852265 (Supplementary Table 5 gives SRA Accession IDs for each sample). These sequences were clustered for Fig. 2c using 'seaborn.clustermap' under default settings, which utilizes the UPGMA algorithm through SciPy[51,52]. The phylogenetic analysis shown in Fig. 3a was produced in Geneious Prime v.2022.1 with PHYML extension, using the GTR substitution model with the optimization conditions of topology/length/rate.

Pathway analysis for SARS-CoV-2 passaged in Huh7-ACE2 cells

Figure 3b was constructed from lineages containing only those mutations found most commonly in passage 16: T21I, T304I, A173V, E166V, P252L, S144A and L50F. These lineages were determined based on the frequencies of the corresponding mutations in a given well at each passage. Pairs of mutants whose frequencies summed to greater than 100% were assumed to co-occur on the same allele. The same logic was extended to identify triple and quadruple mutants, such that if each pairwise sum of frequencies within a group of mutations was greater than 100%, all mutations within that group were assumed to occur together. The order in which mutations in a given lineage arose was imputed either from stepwise appearance over time (for example, passage 4 has mutation 1 and passage 8 has mutations 1 and 2 at a total combined frequency greater than 100%) with increasing frequencies, or, in cases in which two mutations arose between sequenced passages and were deemed to co-occur in a single virus, by their relative frequencies (for example, if passage 4 has no mutations and passage 8 has mutation 1 at 99% frequency and mutation 2 at 30% frequency, mutation 1 was assumed to have arisen first). See Supplementary Table 2 for the datapoints used in this analysis.

Recombinant SARS-CoV-2 production

A reverse genetics system based on the pBeloBAC11 bacterial artificial chromosome (BAC), containing the SARS-CoV-2 genome with a NanoLuc luciferase reporter replacing ORF7a[53] (gift of L.-M. Sobrido), was used to produce recombinant SARS-CoV-2 harbouring $3CL^{pro}$ mutations. Mutant BACs were produced as previously described[33]; see Supplementary Table 6 for a list of mutagenic primers used. These BACs (2 µg each) were then transfected into HEK293T cells in 12-well plates in triplicate using Lipofectamine 3000 Transfection Reagent (ThermoFisher) according to the manufacturer's instructions. Two days post transfection, cells were pooled and overlaid onto Vero E6-TMPRSS2-T2A-ACE2 cells in 25 cm^2 flasks. After 3 days, supernatant was collected from these cells, clarified by centrifugation then used to infect Vero E6 cells in 75 cm^2 flasks. Four days post infection, supernatant was harvested, clarified by centrifugation and aliquoted. Viruses were stored at −80 °C before use. An aliquot of all recombinant viruses was confirmed by nanopore sequencing for the mutation of interest, and for purity before use.

Inhibition assay with recombinant SARS-CoV-2

Viruses were first titrated to normalize input. To characterize inhibition, Huh7-ACE2 cells were seeded at a density of 2×10^4 cells per well in 96-well plates. The following day, cells were infected with 0.05 MOI of virus and treated with inhibitor in a fivefold dilution series. One day post infection, cells were lysed and luminescence quantified using the Nano-Glo Luciferase Assay System (Promega), according to the manufacturer's instructions, with a SpectraMax i3× Multi-Mode Microplate Reader (Molecular Devices) using SoftMax Pro v.7.0.2 software (Molecular Devices). IC_{50} values were derived by fitting a nonlinear regression curve to the data in GraphPad Prism v.9.4 (Dotmatics).

Growth assay with recombinant SARS-CoV-2

Viruses were first titrated to normalize input. To characterize fitness, Huh7-ACE2 cells were seeded at a density of 2×10^4 cells per well in 96-well plates. The following day, cells were infected with 0.01 MOI of virus. At 12, 24, 36 and 48 h post infection, cells were lysed and luminescence quantified

using the Nano-Glo Luciferase Assay System, according to the manufacturer’s instructions, with a SpectraMax i3× Multi-Mode Microplate Reader using SoftMax Pro v.7.0.2 software.

Retrieval of clinical mutation frequencies

COVID-19 CG was used to retrieve all clinically observed 3CL^{pro} mutations from the Global Initiative on Sharing Avian Influenza Data on 26 June 2022, either since the start of the COVID-19 pandemic or for the periods 26 March–26 June 2022 and 22 September–22 December 2021 (refs. [31,54]).

Materials availability

Materials used in this study will be made available under an appropriate materials transfer agreement.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All experimental data are provided in the manuscript. The sequences of mutants from passaging in Vero E6 cells have been deposited at GenBank (nos. ON924329–ON924335 and ON930401–ON930431). The raw NGS data of passaging in Huh7-ACE2 cells are available from the NCBI Sequence Read Archive under BioProject Accession ID PRJNA852265. The structures of the 3CL^{pro}–nirmatrelvir and 3CL^{pro}–ensitrelvir complexes were downloaded from PDB under accession codes 7VH8 and 7VU6, respectively. The Wuhan-Hu-1 sequence used for alignment was downloaded from GenBank (accession no. MN908947).

Code availability

Sequencing data processing and visualization was performed using the ONT ARTIC Nextflow pipeline (https://github.com/epi2me-labs/wf-artic) and the V-pipe computational pipeline (https://github.com/cbg-ethz/V-pipe) and clustering was performed using seaborn (https://github.com/mwaskom/seaborn), which utilizes SciPy (https://github.com/scipy/scipy), all of which are publicly available software packages.

References

1. Hammond, J. et al. Oral nirmatrelvir for high-risk, nonhospitalized adults with Covid-19. *N. Engl. J. Med.* **386**, 1397–1408 (2022).

2. Owen, D. R. et al. An oral SARS-CoV-2 M(pro) inhibitor clinical candidate for the treatment of COVID-19. *Science* **374**, 1586–1593 (2021).

3. Altarawneh, H. N. et al. Effects of previous infection and vaccination on symptomatic omicron infections. *N. Engl. J. Med.* **387**, 21–34 (2022).

4. Iketani, S. et al. Antibody evasion properties of SARS-CoV-2 Omicron sublineages. *Nature* **604**, 553–556 (2022).

5. Liu, L. et al. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022).

6. Planas, D. et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021).

7. Wang, P. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).

8. Wang, Q. et al. Antibody evasion by SARS-CoV-2 Omicron subvariants BA.2.12.1, BA.4, and BA.5. *Nature* **608**, 603–608 (2022).

9. Gandhi, S. et al. De novo emergence of a remdesivir resistance mutation during treatment of persistent SARS-CoV-2 infection in an

immunocompromised patient: a case report. *Nat. Commun.* **13**, 1547 (2022).

10. Baden, L. R. et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.* **384**, 403–416 (2021).

11. Chen, P. et al. SARS-CoV-2 neutralizing antibody LY-CoV555 in outpatients with Covid-19. *N. Engl. J. Med.* **384**, 229–237 (2021).

12. Dougan, M. et al. Bebtelovimab, alone or together with bamlanivimab and etesevimab, as a broadly neutralizing monoclonal antibody treatment for mild to moderate, ambulatory COVID-19. Preprint at *medRxiv* https://doi.org/10.1101/2022.03.10.22272100 (2022).

13. Gupta, A. et al. Early treatment for Covid-19 with SARS-CoV-2 neutralizing antibody sotrovimab. *N. Engl. J. Med.* **385**, 1941–1950 (2021).

14. Polack, F. P. et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N. Engl. J. Med.* **383**, 2603–2615 (2020).

15. Sadoff, J. et al. Safety and efficacy of single-dose Ad26.COV2.S vaccine against Covid-19. *N. Engl. J. Med.* **384**, 2187–2201 (2021).

16. Weinreich, D. M. et al. REGN-COV2, a neutralizing antibody cocktail, in outpatients with Covid-19. *N. Engl. J. Med.* **384**, 238–251 (2021).

17. Beigel, J. H. et al. Remdesivir for the treatment of Covid-19 – final report. *N. Engl. J. Med.* **383**, 1813–1826 (2020).

18. Gottlieb, R. L. et al. Early remdesivir to prevent progression to severe Covid-19 in outpatients. *N. Engl. J. Med.* **386**, 305–315 (2022).

19. Fischer, W. A. 2nd et al. A phase 2a clinical trial of molnupiravir in patients with COVID-19 shows accelerated SARS-CoV-2 RNA clearance and elimination of infectious virus. *Sci. Transl. Med.* **14**, eabl7430 (2022).

20. Jayk Bernal, A. et al. Molnupiravir for oral treatment of Covid-19 in nonhospitalized patients. *N. Engl. J. Med.* **386**, 509–520 (2022).

21. Painter, W. P. et al. Human safety, tolerability, and pharmacokinetics of molnupiravir, a novel broad-spectrum oral antiviral agent with activity against SARS-CoV-2. *Antimicrob. Agents Chemother.* **65**, e02428–20 (2021).

22. Chan, A. P., Choi, Y. & Schork, N. J. Conserved genomic terminals of SARS-CoV-2 as coevolving functional elements and potential therapeutic targets. *mSphere* **5**, e00754–20 (2020).

23. V'Kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **19**, 155–170 (2021).

24. Stevens, L. J. et al. Mutations in the SARS-CoV-2 RNA dependent RNA polymerase confer resistance to remdesivir by distinct mechanisms. *Sci. Transl. Med.* **14**, eabo0718 (2022).

25. Szemiel, A. M. et al. In vitro selection of Remdesivir resistance suggests evolutionary predictability of SARS-CoV-2. *PLoS Pathog.* **17**, e1009929 (2021).

26. Hoffman, R. L. et al. Discovery of ketone-based covalent inhibitors of coronavirus 3CL proteases for the potential therapeutic treatment of COVID-19. *J. Med. Chem.* **63**, 12725–12747 (2020).

27. Xie, X. et al. An infectious cDNA clone of SARS-CoV-2. *Cell Host Microbe* **27**, 841–848 (2020).

28. Unoh, Y. et al. Discovery of S-217622, a noncovalent oral SARS-CoV-2 3CL protease inhibitor clinical candidate for treating COVID-19. *J. Med. Chem.* **65**, 6499–6512 (2022).

29. Mukae, H. et al. A randomized phase 2/3 study of ensitrelvir, a novel oral SARS-CoV-2 3C-like protease inhibitor, in Japanese patients with mild-to-moderate COVID-19 or asymptomatic SARS-CoV-2 infection:

results of the phase 2a part. *Antimicrob. Agents Chemother*. **66**, e0069722 (2022).

30. Ferreira, J. C., Fadl, S. & Rabeh, W. M. Key dimer interface residues impact the catalytic activity of 3CLpro, the main protease of SARS-CoV-2. *J. Biol. Chem.* **298**, 102023 (2022).

31. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro. Surveill.* **22**, 30494 (2017).

32. ASPR. *COVID-19 Therapeutics Thresholds, Orders, and Replenishment by Jurisdiction* https://aspr.hhs.gov/COVID-19/Therapeutics/orders/Pages/default.aspx (2022).

33. Iketani, S. et al. Functional map of SARS-CoV-2 3CL protease reveals tolerant and immutable sites. *Cell Host Microbe* **30**, 1354–1362.e6 (2022).

34. Jochmans, D. et al. The substitutions L50F, E166A and L167F in SARS-CoV-2 3CLpro are selected by a protease inhibitor in vitro and confer resistance to nirmatrelvir. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.07.495116 (2022).

35. Zhou, Y. et al. Nirmatrelvir resistant SARS-CoV-2 variants with high fitness in vitro. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.06.494921 (2022).

36. Xue, X. et al. Structures of two coronavirus main proteases: implications for substrate binding and antiviral drug design. *J. Virol.* **82**, 2515–2527 (2008).

37. Yang, S. et al. Synthesis, crystal structure, structure-activity relationships, and antiviral activity of a potent SARS coronavirus 3CL protease inhibitor. *J. Med. Chem.* **49**, 4971–4980 (2006).

38. Zhao, Y. et al. Crystal structure of SARS-CoV-2 main protease in complex with protease inhibitor PF-07321332. *Protein Cell* **13**, 689–693 (2022).

39. US Food and Drug Administration. Fact sheet for healthcare providers: emergency use authorization for Paxlovid™. https://www.fda.gov/media/155050/download (accessed 20 July 2022).

40. Padhi, A. K. & Tripathi, T. Hotspot residues and resistance mutations in the nirmatrelvir-binding site of SARS-CoV-2 main protease: design, identification, and correlation with globally circulating viral genomes. *Biochem. Biophys. Res. Commun.* **629**, 54–60 (2022).

41. Shaqra, A. M. et al. Defining the substrate envelope of SARS-CoV-2 main protease to predict and avoid drug resistance. *Nat. Commun.* **13**, 3556 (2022).

42. Yang, K. S., Leeuwon, S. Z., Xu, S. & Liu, W. R. Evolutionary and structural insights about potential SARS-CoV-2 evasion of nirmatrelvir. *J. Med. Chem.* **65**, 8686–8698 (2022).

43. Sacco, M. D. et al. The P132H mutation in the main protease of Omicron SARS-CoV-2 decreases thermal stability without compromising catalysis or small-molecule drug inhibition. *Cell Res.* **32**, 498–500 (2022).

44. Liu, H. et al. Development of optimized drug-like small molecule inhibitors of the SARS-CoV-2 3CL protease for treatment of COVID-19. *Nat. Commun.* **13**, 1891 (2022).

45. Liu, L. et al. Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450–456 (2020).

46. Freed, N. E., Vlkova, M., Faisal, M. B. & Silander, O. K. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol. Methods Protoc.* **5**, bpaa014 (2020).

47. Annavajhala, M. K. et al. Emergence and expansion of SARS-CoV-2 B.1.526 after identification in New York. *Nature* **597**, 703–708 (2021).

48. Reed, L. J. & Muench, H. A simple method of estimating fifty per cent endpoints. *Am. J. Epidemiol.* **27**, 493–497 (1938).

49. Stirling, D. R. et al. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 433 (2021).

50. Posada-Cespedes, S. et al. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics* **37**, 1673–1680 (2021).

51. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

52. Waskom, M. L. seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

53. Ye, C. et al. Rescue of SARS-CoV-2 from a single bacterial artificial chromosome. *mBio* **11**, e02168–20 (2020).

54. Chen, A. T., Altschuler, K., Zhan, S. H., Chan, Y. A. & Deverman, B. E. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *eLife* **10**, e63409 (2021).

Acknowledgements

This study was supported by funding from the JPB Foundation, A. and P. Cherng, S. Yin, C. Ludwig and D. and R. Wu to D.D.H. A.C. is supported by a Career Awards for Medical Scientists from the Burroughs Wellcome Fund. We thank P.-Y. Shi for the SARS-CoV-2-mNeonGreen reporter virus and L. Martinez-Sobrido and C. Ye for the bacterial artificial chromosome system used to generate recombinant SARS-CoV-2.

Author information

Author notes

1. These authors contributed equally: Sho Iketani, Hiroshi Mohri, Bruce Culbertson, Seo Jung Hong

Authors and Affiliations

1. Aaron Diamond AIDS Research Center, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

   Sho Iketani, Hiroshi Mohri, Maria I. Luck, Yicheng Guo, Zizhang Sheng, Stephen P. Goff, Yosef Sabo & David D. Ho

2. Division of Infectious Diseases, Department of Medicine, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

   Sho Iketani, Hiroshi Mohri, Maria I. Luck, Medini K. Annavajhala, Yicheng Guo, Zizhang Sheng, Anne-Catrin Uhlemann, Yosef Sabo & David D. Ho

3. Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

   Bruce Culbertson

4. Medical Scientist Training Program, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

   Bruce Culbertson

5. Department of Pathology and Cell Biology, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

   Seo Jung Hong & Alejandro Chavez

6. Shanghai Institute for Advanced Immunochemical Studies and School of Life Science and Technology, ShanghaiTech University, Shanghai, China

   Yinkai Duan & Haitao Yang

7. Department of Microbiology and Immunology, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

Stephen P. Goff & David D. Ho

8. Department of Biochemistry and Molecular Biophysics, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

Stephen P. Goff

Contributions

S.I., B.C., S.J.H., A.C. and D.D.H. conceived the project and the approach to scaled screening for resistant viral variants. S.I. and H.M. conducted in vitro passaging. S.I., H.M., B.C., S.J.H., M.K.A., A.-C.U. and A.C. conducted sequencing. S.I. and H.M. conducted fitness and inhibition assays. B.C., S.J.H. and A.C. conducted pathway analyses. S.I., H.M., B.C., S.J.H., M.I.L., Y.S. and A.C. generated recombinant SARS-CoV-2. Y.D., Y.G., Z.S. and H.Y. conducted structural analyses. S.P.G. contributed to discussions of the data and analysis. S.I., H.M., B.C., S.J.H., Y.D., H.Y., A.C. and D.D.H. wrote the manuscript with input from all authors.

Corresponding authors

Correspondence to Alejandro Chavez or David D. Ho.

Ethics declarations

Competing interests

S.I., A.C. and D.D.H. are inventors on patent applications related to the development of inhibitors against the SARS-CoV-2 3CL protease. D.D.H. is a cofounder of TaiMed Biologics and RenBio, consultant to WuXi Biologics and Brii Biosciences and board director for Vicarious Surgical.

Peer review

Peer review information

Nature thanks Rolf Hilgenfeld and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Growth assays with SARS-CoV-2 passaged in Vero E6 cells.

a-**d**, Growth was quantified for lineage A (**a**), lineage B (**b**), lineage C (**c**), and unpassaged SARS-CoV-2 (**d**, denoted as WT-P0) in comparison to SARS-CoV-2 passaged without nirmatrelvir for 30 passages (denoted as WT-P30). Vero E6 cells were infected with 200 $TCID_{50}$ of the indicated viruses and viral RNA was quantified at the indicated time points. **e**, The slope during the exponential phase (between 11 and 24 h post-infection) of growth for the indicated viruses.

Extended Data Fig. 2 Mutations in the 11 $3CL^{pro}$ cut sites found in passage 16 from the 53 wells passaged in Huh7-ACE2 cells.

Dots indicate wild-type at that cut site. Note that nsp4/5 M(P6')I = M6I, nsp5/6 S(P6)P = S301P, and nsp5/6 T(P3)I = T304I in $3CL^{pro}$.

Extended Data Fig. 3 Mutations studied as isogenic recombinant SARS-CoV-2 overlaid onto the 3CL protease structure.

The Cα of each residue that was mutated is denoted with a red sphere. The $3CL^{pro}$-nirmatrelvir complex was downloaded from PDB under accession code 7VH8.

Extended Data Fig. 4 Raw IC_{50} values for recombinant live SARS-CoV-2 carrying single and combination $3CL^{pro}$ mutations by nirmatrelvir, ensitrelvir, and remdesivir.

Mean ± SD of three biologically independent experiments are shown.

Extended Data Fig. 5 Individual inhibition curves of recombinant live SARS-CoV-2 carrying single and combination $3CL^{pro}$ mutations by ensitrelvir and remdesivir.

Representative curves from a single experiment from three biologically independent experiments are shown. Error bars denote mean ± s.e.m of three technical replicates.

Extended Data Fig. 6 Inhibition of passage 30 of SARS-CoV-2 passaged in Vero E6 cells by nirmatrelvir, ensitrelvir, and remdesivir.

a, Raw IC_{50} values. **b**, Fold change relative to inhibition of wild-type.

Extended Data Fig. 7 Structural analyses of $3CL^{pro}$ mutations.

a, Overlay of nirmatrelvir and ensitrelvir binding to $3CL^{pro}$. **b**, Several of the residues involved in direct interaction with nirmatrelvir. **c**, Several of the residues involved in formation of the S1 subsite. **d**, Interaction of L167 with nirmatrelvir. In **a**-**d**, nirmatrelvir is shown in yellow, enstirelvir is shown in lime green, the $3CL^{pro}$-nirmatrelvir complex is shown in marine, and the $3CL^{pro}$-ensitrelvir complex is shown in gray. Protomer A is shown in marine and protomer B is shown in green. Hydrogen bonds are indicated as black dashes. The $3CL^{pro}$-nirmatrelvir complex and $3CL^{pro}$-ensitrelvir complex were downloaded from PDB under accession codes 7VH8 and 7VU6, respectively.

Extended Data Fig. 8 Frequencies of identified 3CL^{pro} mutations in GISAID.

a, All occurrences of the indicated mutations were tabulated from GISAID. **b**, All occurrences of the indicated mutations were tabulated from GISAID in the three months prior to EUA (9/22/2021 to 12/22/2021) or after EUA (3/26/2022 to 6/26/2022).

Supplementary information

Reporting Summary

Supplementary Table 1

Raw sequencing results for passaging in Huh7-ACE2 cells. Samples are denoted as passage number, followed by well number. Cut site mutations are denoted as 'cs'.

Supplementary Table 2

Passage transitions used for construction of the pathway analysis in Fig. 3b.

Supplementary Table 3

GenBank Accession IDs for sequences from passaging in Vero E6 cells.

Supplementary Table 4

Oligos used for NGS.

Supplementary Table 5

SRA Accession IDs for raw sequencing data from passaging in Huh7-ACE2 cells.

[Supplementary Table 6]

Oligos used for site-directed mutagenesis to produce isogenic recombinant SARS-CoV-2.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Iketani, S., Mohri, H., Culbertson, B. *et al.* Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir. *Nature* **613**, 558–564 (2023). https://doi.org/10.1038/s41586-022-05514-2

- Received: 06 July 2022
- Accepted: 02 November 2022
- Published: 09 November 2022
- Issue Date: 19 January 2023

- DOI: https://doi.org/10.1038/s41586-022-05514-2

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

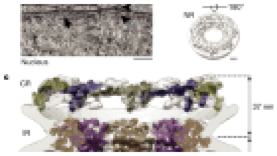
Provided by the Springer Nature SharedIt content-sharing initiative

This article is cited by

- **A strategy for evaluating potential antiviral resistance to small molecule drugs and application to SARS-CoV-2**
    - Karen Sargsyan
    - Karine Mazmanian
    - Carmay Lim

  Scientific Reports (2023)

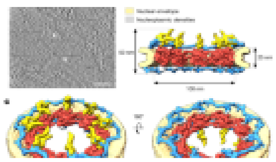
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05514-2



The cellular environment shapes the nuclear pore complex architecture

13 October 2021

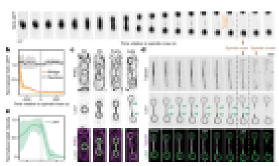
Anthony P. Schuller, Matthias Wojtynek, … Thomas U. Schwartz



In-cell architecture of the nuclear pore and snapshots of its turnover

02 September 2020

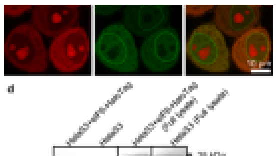
Matteo Allegretti, Christian E. Zimmerli, … Martin Beck



Closed mitosis requires local disassembly of the nuclear envelope

26 August 2020

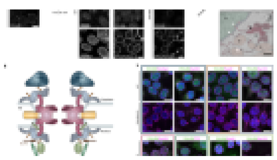
Gautam Dey, Siân Culley, … Buzz Baum



[Nuclear export of the pre-60S ribosomal subunit through single nuclear pores observed in real time]

27 October 2021

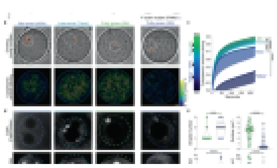
Jan Andreas Ruland, Annika Marie Krüger, … Ulrich Kubitscheck



[The chaperone DNAJB6 surveils FG-nucleoporins and is required for interphase nuclear pore complex biogenesis]

27 October 2022

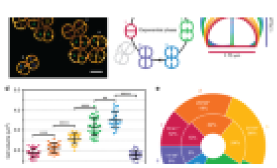
E. F. Elsiena Kuiper, Paola Gallardo, … Steven Bergink



[Cytoplasmic forces functionally reorganize nuclear condensates in oocytes]

29 August 2022

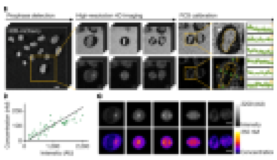
Adel Al Jord, Gaëlle Letort, … Marie-Hélène Verlhac



[Cell morphology and nucleoid dynamics in dividing Deinococcus radiodurans]

23 August 2019

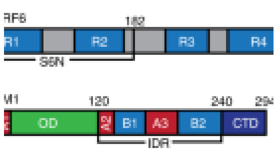
Kevin Floc'h, Françoise Lacroix, … Joanna Timmins



[Experimental and computational framework for a dynamic protein atlas of human cell division]

10 September 2018

Yin Cai, M. Julius Hossain, … Jan Ellenberg



[Compositional adaptability in NPM1-SURF6 scaffolding networks enabled by dynamic switching of phase separation mechanisms]

29 November 2018

Mylene C. Ferrolino, Diana M. Mitrea, … Richard W. Kriwacki

- Article
- Open Access
- Published: 04 January 2023

A quantitative map of nuclear pore assembly reveals two distinct mechanisms

- Shotaro Otsuka ,
- Jeremy O. B. Tempkin,
- Wanlu Zhang,
- Antonio Z. Politi,
- Arina Rybina,
- M. Julius Hossain,
- Moritz Kueblbeck,
- Andrea Callegari,
- Birgit Koch,
- Natalia Rosalia Morero,
- Andrej Sali &
- …
- Jan Ellenberg

Show authors

Nature **volume 613**, pages 575–581 (2023)

- 7724 Accesses

- 55 Altmetric

- Metrics details

Abstract

Understanding how the nuclear pore complex (NPC) is assembled is of fundamental importance to grasp the mechanisms behind its essential function and understand its role during the evolution of eukaryotes[1,2,3,4]. There are at least two NPC assembly pathways—one during the exit from

mitosis and one during nuclear growth in interphase—but we currently lack a quantitative map of these events. Here we use fluorescence correlation spectroscopy calibrated live imaging of endogenously fluorescently tagged nucleoporins to map the changes in the composition and stoichiometry of seven major modules of the human NPC during its assembly in single dividing cells. This systematic quantitative map reveals that the two assembly pathways have distinct molecular mechanisms, in which the order of addition of two large structural components, the central ring complex and nuclear filaments are inverted. The dynamic stoichiometry data was integrated to create a spatiotemporal model of the NPC assembly pathway and predict the structures of postmitotic NPC assembly intermediates.

Main

The nuclear pore complex (NPC) is the largest non-polymeric protein complex in eukaryotic cells. It spans the double membrane of the nucleus—the nuclear envelope—and mediates macromolecular transport between the nucleus and the cytoplasm. To achieve this essential function, the NPC forms an octameric proteinaceous channel composed of multiples of 8 of more than 30 different nucleoporins (Nups) that form 6–8 protein modules, the NPC subcomplexes[1,2]. Therefore, more than 500 individual proteins have to come together to assemble one nuclear pore, which has the mass of tens of ribosomes. NPCs are thought to represent a key step in the evolution of endomembrane compartmentalization that allowed ancestral eukaryotes to separate their genome from the cytoplasm[3,4].

In proliferating cells that undergo open mitosis, there are two main pathways by which NPCs assemble. During nuclear assembly after mitosis, NPCs form together with nuclear membranes to rapidly build new nuclei in the daughter cells—this process is called postmitotic NPC assembly. During nuclear growth in interphase, NPCs continue to assemble continuously for homeostasis—this is referred to as interphase assembly. Research over the past decade has revealed that postmitotic and interphase NPC assembly possess distinct kinetic, molecular and structural features[5,6,7,8,9,10,11,12], suggesting that two fundamentally different mechanisms build the same protein complex. In postmitotic assembly, several thousand NPCs assemble within a few minutes during sealing of the initially fenestrated nuclear

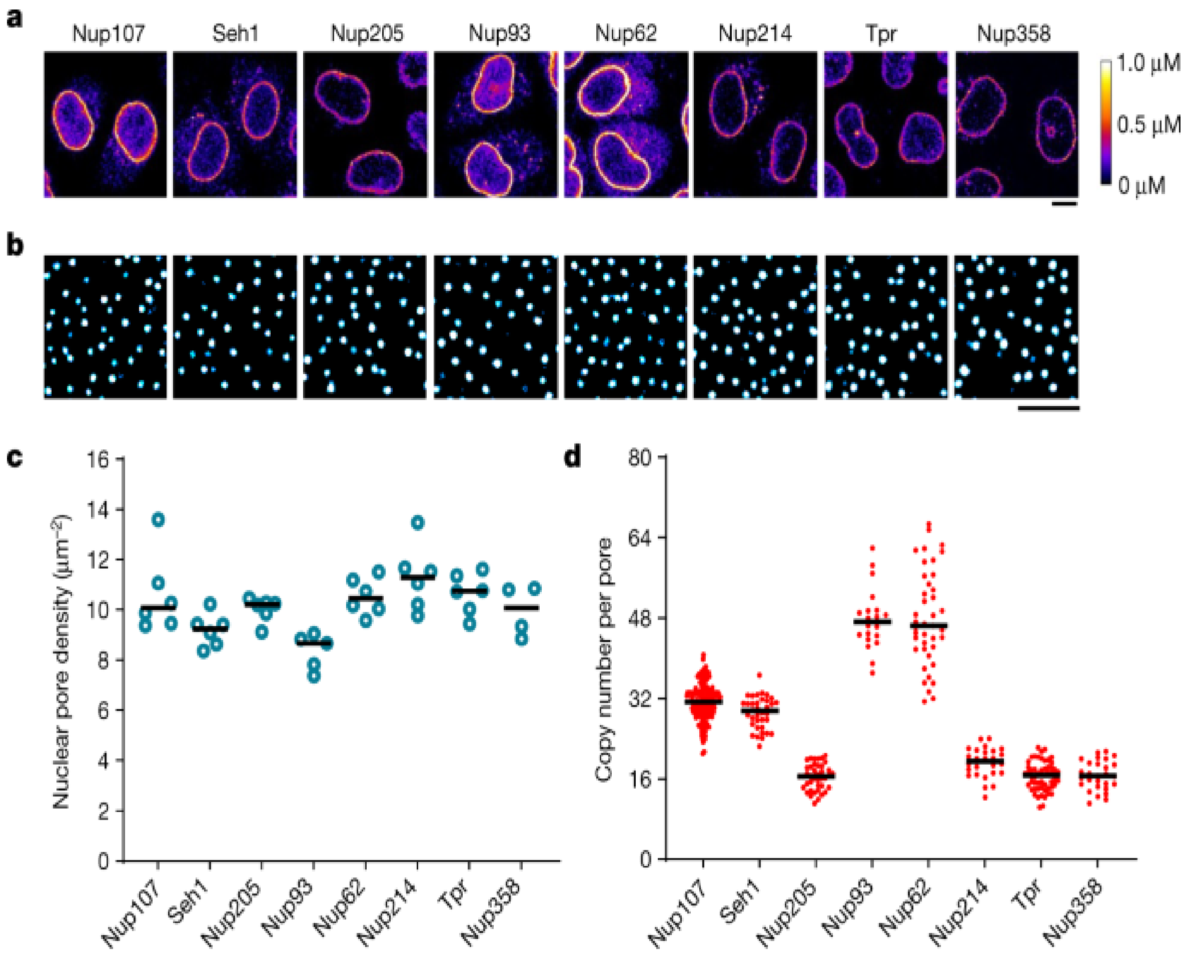
membranes, whereas interphase NPC assembly occurs more sporadically, requires about one hour, and involves a new discontinuity in the double membrane barrier of the nuclear envelope. Studies using molecular depletions have shown that the Nup ELYS is required for postmitotic assembly but appears dispensable for interphase assembly[5], whereas the membrane curvature-sensing domain of Nup133[6], Pom121 and Sun1[7,8], the import of Nup153 into the nucleus[9], and Torsins[10] seem to be required only for interphase assembly. Recent studies correlating real-time imaging with 3D electron microscopy have revealed that postmitotic NPC assembly proceeds by radial dilation of small membrane openings[11], whereas in interphase, assembly induces an asymmetric inside-out fusion of the inner and outer nuclear membranes[12].

However, how several hundred proteins self-organize to form the NPC channel via these two distinct assembly pathways has remained largely unknown. It is technically challenging to locate the transient and rare assembly events, which has prevented investigation of the structure of assembly intermediates by either cryo-electron microscopy (cryo-EM) tomography or super-resolution microscopy. In addition, the large number of building blocks and their cooperativity often leads to complex nonlinear kinetics that can only be interpreted mechanistically using computational modelling of the structures formed during assembly. Although we have some information about the dynamic addition of Nups after mitosis[13,14], only sparse dynamic data is available for interphase assembly[15,16]. Notably, these earlier studies could not distinguish between postmitotic and interphase assemblies, whose co-occurrence in different regions of the nucleus was only discovered later; moreover, these studies provided only qualitative descriptions, as they relied on ectopic expression of fluorescently tagged Nups. Kinetic data about NPC assembly that can distinguish between the postmitotic and interphase pathways is required, including the copy numbers of Nups that assemble into forming NPCs over time. Such data would enable modelling of the assembly process and enable us to understand the two assembly mechanisms.

Quantitative imaging of ten Nups

To quantitatively analyse the changes in concentration of Nups at the nuclear envelope during exit from mitosis and nuclear growth in G1, we genome-edited HeLa cells, homozygously tagging the endogenous genes encoding different Nups with mEGFP or mCherry. Building on previous work[11,12,17,18], we created a set of ten endogenously tagged Nups that systematically represent the major building blocks of the fully assembled pore, including the nuclear filament protein Tpr, the nuclear Nup153, the Y-complex members Nup107 and Seh1, the central ring complex members Nup93 and Nup205, the central channel protein Nup62, the transmembrane protein Pom121, as well as the cytoplasmic filament proteins Nup214 and Nup358. Homozygous tagging was verified by careful quality control of the genome-edited monoclonal cell lines[19], ensuring that the tagged subunit was localized to the NPC and that cell viability and mitotic progression were normal (Figs. 1 and 2). Tagged Nup153, Pom121 and Nup358 were expressed at subphysiological levels (Extended Data Figs. 1 and 2; details in Discussion). The fusion proteins are likely to be functional, given that most Nups show strong phenotypes upon knockout or depletion[20].

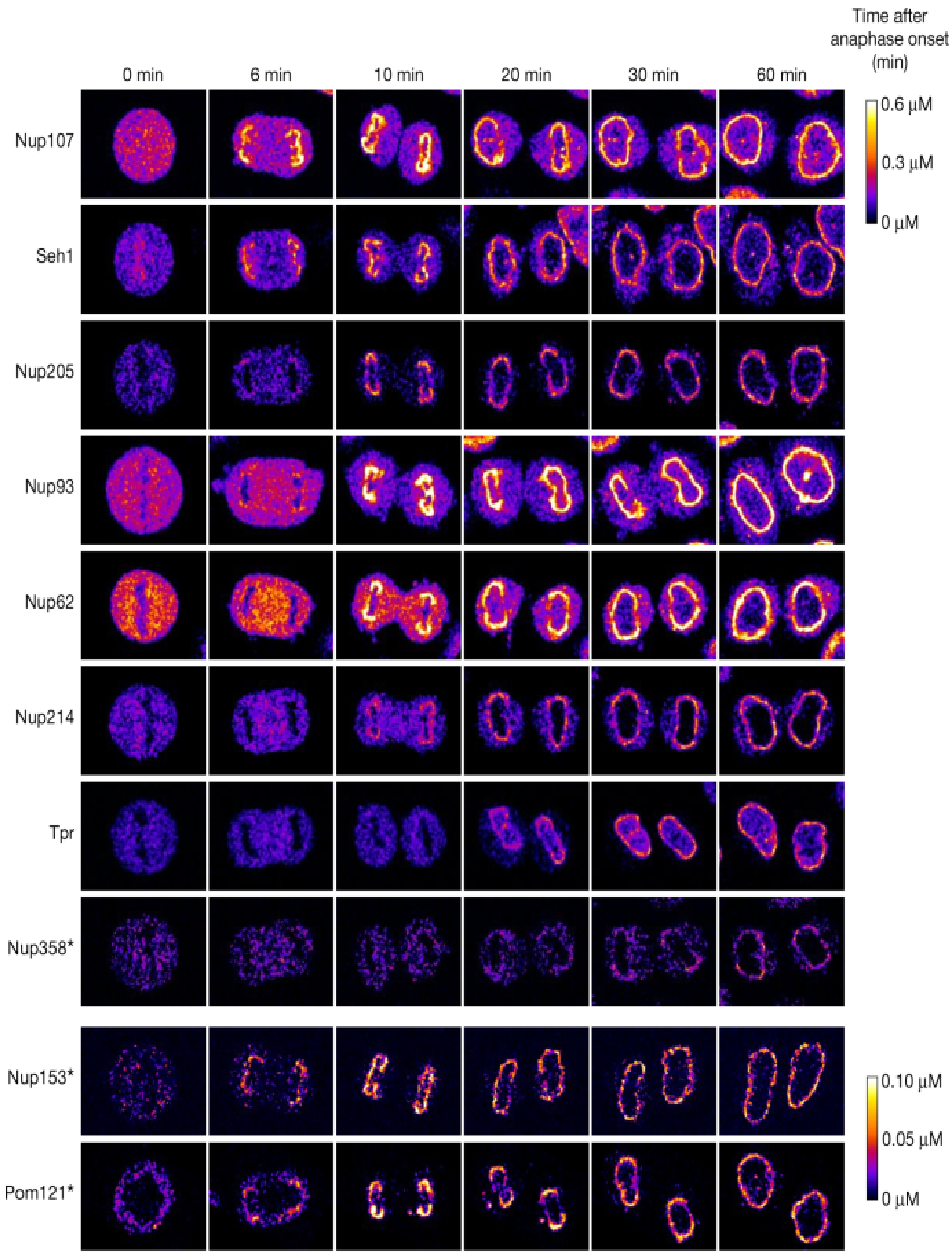
Fig. 1: Quantitative imaging of GFP-knock-in Nup cell lines.



a, Confocal microscopy of genome-edited HeLa cells with homozygous expression of mEGFP-tagged Nups. Fluorescence intensity was converted to protein concentration (colour bar) using FCS-calibrated imaging[22]. Images were filtered with a median filter (kernel size: 0.25 × 0.25 µm) for presentation purposes. Scale bar, 10 µm. **b**,**c**, Stimulated emission depletion (STED) microscopy of genome-edited cells stained with Nup62 antibody. Cells were imaged (**b**) and the density of nuclear pores was quantified (**c**). *n* = 6 (Nup107), 6 (Seh1), 6 (Nup205), 5 (Nup93), 6 (Nup62), 6 (Nup214), 6 (Tpr) and 4 (Nup358) cells. Scale bar, 1 µm. **d**, Calculated copy number of Nups per nuclear pore. *n* = 241 (Nup107), 37 (Seh1), 41 (Nup205), 20 (Nup93), 41 (Nup62), 26 (Nup214), 55 (Tpr) and 28 (Nup358) cells. The horizontal line represents the median.

Source data

Fig. 2: Dynamic concentration maps of Nups after anaphase onset.



HeLa cells whose Nups are endogenously tagged with mEGFP or mCherry were imaged every 30 s by 3D confocal microscopy. Single confocal

sections are shown. Images were calibrated by FCS to convert fluorescence intensities into cellular protein concentration (colour bars). Nup153 is heterozygously tagged and Nup358 and Pom121 are subphysiologically expressed (Extended Data Figs. 1 and 2). The bottom colour bar is for Nup153 and Pom121; the top colour bar is for the rest of the Nups. Images were filtered with a median filter (kernel size: 0.25 × 0.25 µm). Scale bar, 10 µm.

To characterize the fluorescently tagged Nups, we first performed super-resolution (STED) microscopy to determine the NPC density in fully grown nuclei of the knock-in cell lines, showing that homozygous tagging had little effect on NPC density, which was within a 15% range between all cell lines, with an average of 10.1 NPCs per μm^2 (Fig. 1b,c), in good agreement with our previous estimates from electron microscopy studies on HeLa cells[12]. We then used fluorescence correlation spectroscopy (FCS) calibrated confocal microscopy[21,22] to determine the concentration and total number of Nups at the nuclear envelope in living cells (Supplementary Table 1; details in Methods). Using the measured NPC density and the Nup concentration at the nuclear envelope, we were able to calculate the average copy number of each Nup per NPC (Fig. 1d). These data showed that the investigated Nups were present in 16, 32, or 48 copies per pore on average, as expected from the eight-fold symmetry of the complex and are in good overall agreement with previous estimates from mass spectrometry[23].

Dynamic change of Nup numbers

We then used our cell line resource to quantitatively image the Nups during both postmitotic and interphase NPC assembly, from metaphase until the end of the rapid nuclear growth phase in G1, 2 h after anaphase onset. To this end, we performed systematic FCS-calibrated 3D confocal time-lapse microscopy[17] (Fig. 2). Using the single-molecule fluctuation calibration, the 4D imaging data could be converted into maps of subcellular protein concentration (Fig. 2). Counterstaining live nuclei with silicon–rhodamine (SiR)–DNA[24] enabled computational image segmentation to measure the soluble cytoplasmic pool and the nuclear envelope-associated pool over time[17]. Temporal alignment to anaphase onset then enabled us to compare

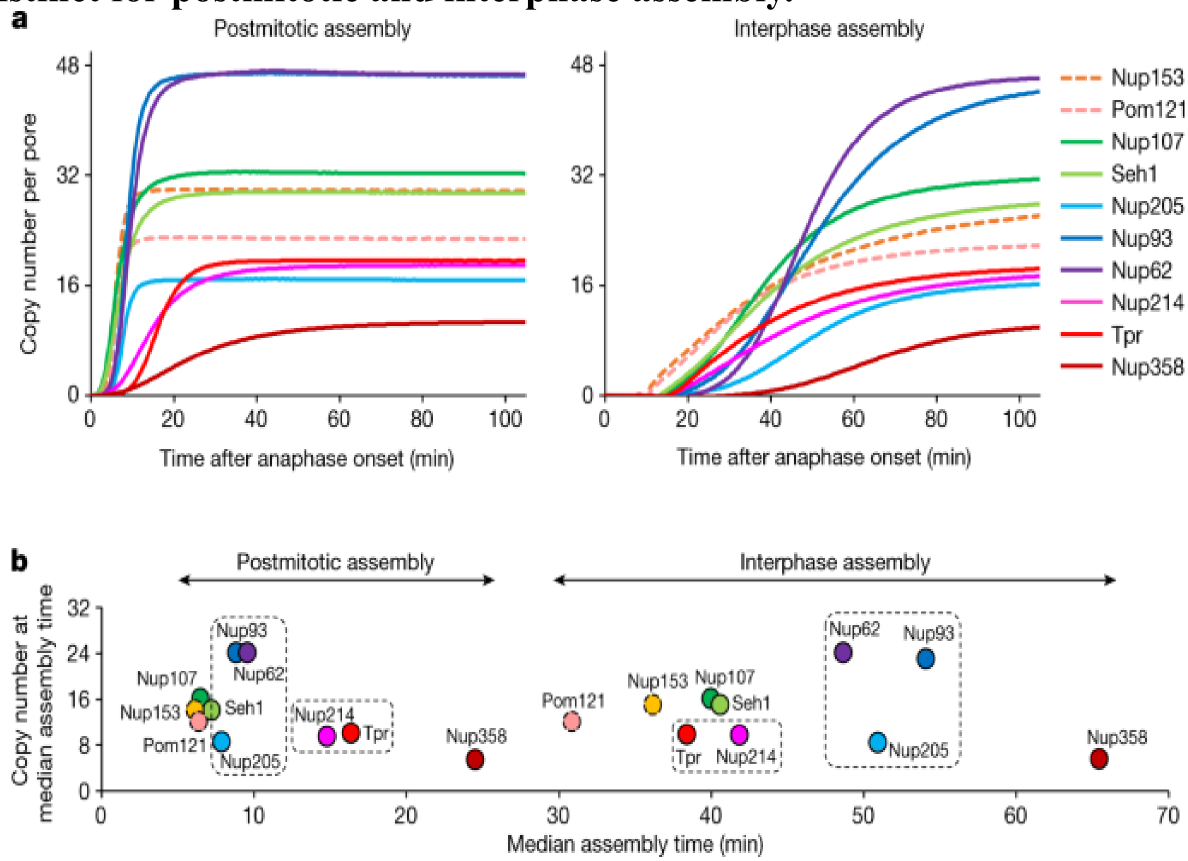
the dynamic association of all Nups with the nuclear envelope over time (Fig. 2). Overall, the investigated Nups are present in 250,000 to 1,200,000 copies per human metaphase cell. After mitosis, this building material is split between the daughter cells with little detectable new protein synthesis in the first hour after anaphase onset (Extended Data Fig. 3). Notably, 34–53% of the soluble pool present in the cytoplasm in metaphase was rapidly re-localized to the nuclear envelope during the first hour after exit from mitosis (43, 38, 47, 37, 41, 44, 53 and 34% for Nup107, Seh1, Nup205, Nup93, Nup62, Nup214, Tpr and Nup358, respectively) (Extended Data Fig. 3), indicating that NPC assembly initially relies almost entirely on the pool of building blocks inherited from the mother cell to form the first 4,000 to 5,000 NPCs[11,12]. We confirmed that the observed GFP-tagged Nups are indeed recruited specifically to NPCs and not to the nuclear envelope surface in general by imaging of single NPCs using STED microscopy (Extended Data Fig. 4a,b).

Quantification of assembly kinetics

Postmitotic and interphase NPC assembly can be observed in the same living cell in different regions of the nuclear envelope within the first 2 h after mitosis[12]. Whereas postmitotic assembly dominates the peripheral 'non-core' regions of the nuclear envelope, the central 'inner-core' area is populated with NPCs only after exit from mitosis[25], when dense spindle microtubules have been removed from the DNA surface[26]. Using computational segmentation and assignment of the inner-core and non-core regions (Extended Data Fig. 5a; details in Methods, modified from ref. [12]), we measured the changes in concentration of the ten Nups in these two regions separately. A two-component model of a fast (postmitotic) and a slow (interphase) assembly process fits the experimental data well, enabling us to kinetically unmix the two assembly processes for each Nup (Extended Data Figs. 5b–d and 6). In this way, we were able to perform an integrated analysis of the real-time kinetics of absolute concentration changes of Nups in all major NPC modules during the two assembly processes at the nuclear envelope (Fig. 3a). This analysis immediately revealed that the overall duration of the two processes is very different, with postmitotic assembly essentially complete 15 min after anaphase onset, whereas interphase

assembly only reaches a plateau after 100 min, consistent with our previous estimates based on live-cell imaging[13,16] and correlative electron microscopy[11,12]. Both processes reached the same final ratios between the different Nups and thus presumably formed identical NPCs. However, the temporal orders in which components were added were distinct, including—for example—an earlier assembly of Pom121 relative to the Y-shaped complex during interphase assembly, consistent with our previous observations[16].

Fig. 3: The molecular assembly order and maturation kinetics are distinct for postmitotic and interphase assembly.



a, The average copy number per nuclear pore computed from mathematical modelling for postmitotic (left) and interphase (right) assembly (details in Methods and Extended Data Figs. 5 and 6). For Nup153 and Pom121 (dashed lines), the absolute amount was estimated using the copy number determined from a previous study[23] (32 for Nup153 and 16 for Pom121). **b**, The average copy number of individual Nups per nuclear pore are plotted along their median assembly time in postmitotic and interphase assembly

pathways. Boxes highlight the Nups that show marked differences in their order of assembly between the two pathways.

Source data

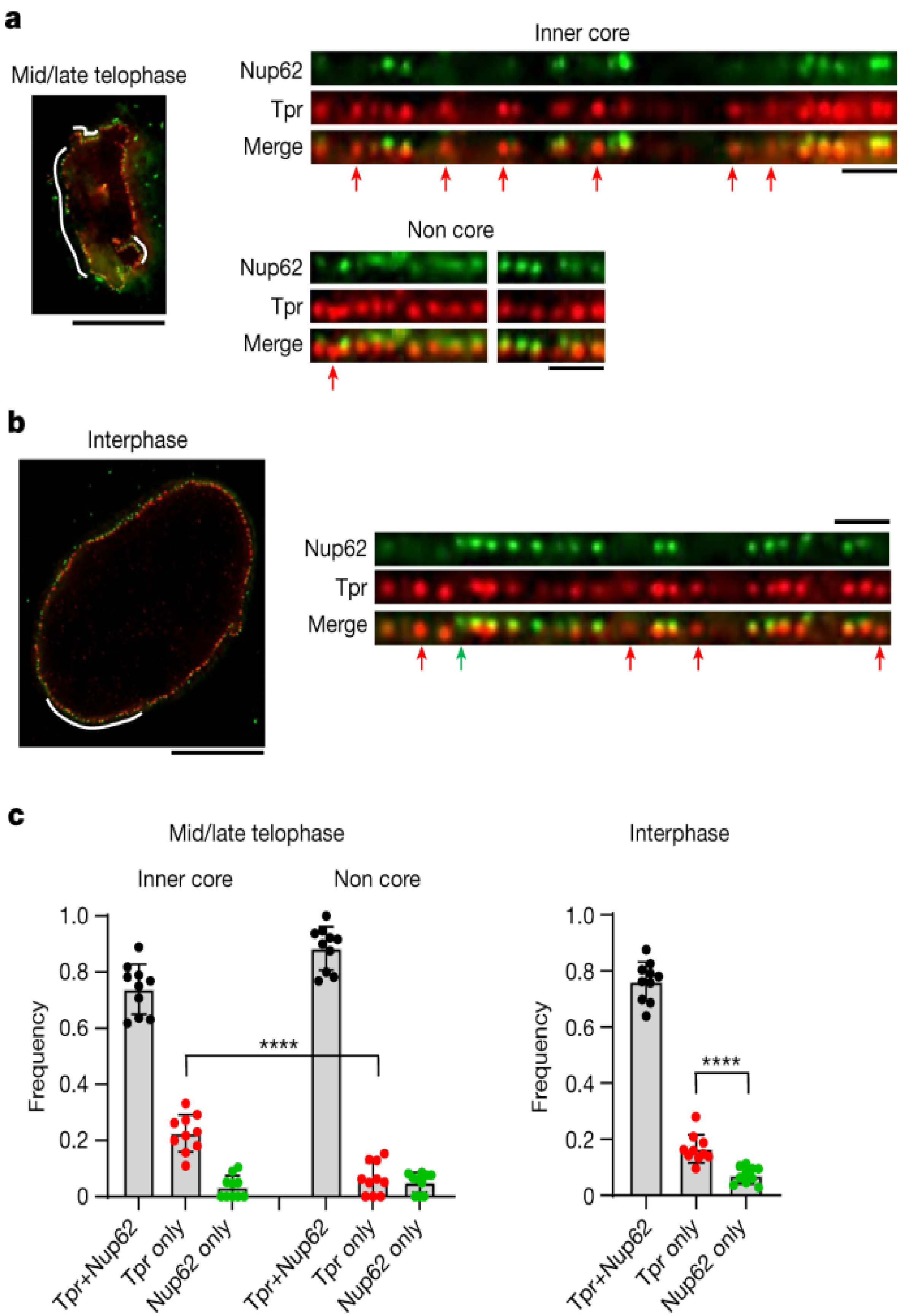
To comprehensively investigate the molecular differences between the two assembly processes, we relied on the constant NPC density and changes in nuclear surface area[12] to convert the nuclear envelope concentrations of all investigated Nups into their average copy number per assembling NPC over time (Fig. 3a). This result in turn enabled us to estimate changes in subunit stoichiometry of the complex during its assembly in living cells. To facilitate the comparative analysis of the assembly kinetics between the two pathways, in which ten components assemble with different speeds and orders, we first reduced the dimensionality of the kinetic data. To this end, we assigned a single characteristic time point of assembly to each Nup by sigmoidal fitting of its full kinetic signature (Extended Data Fig. 7a). Plotting the copy number against the median assembly time provides an overview of the major molecular differences between the two assembly pathways (Fig. 3b).

Whereas the Y-complex, Pom121 and Nup153 form a core of the first modules that assemble almost simultaneously within one minute in postmitotic assembly, these components are stretched out into a clear temporal order of first Nup153, second Pom121 and third the Y-complex (notably with its two investigated subunits also assembling simultaneously in interphase) over more than ten minutes in interphase assembly. The end of assembly is marked for both processes by the addition of the large cytoplasmic filament protein Nup358. The major difference is thus not in initiation or termination of assembly, but rather in the middle of the two assembly pathways. During postmitotic assembly, the Y-complex is rapidly combined with components of the central ring, building the inner core of the pore within just three minutes before addition of either cytoplasmic or nuclear filament proteins, which follow later. By contrast, during interphase assembly, the Y-complex is first combined with the nuclear filament protein Tpr and the base of the cytoplasmic filament Nup214, and the central ring complex is added later. This observation clearly shows that postmitotic and interphase NPC assembly proceed with different speeds and follow different

molecular mechanisms, with inverted molecular orders between the central ring and nuclear filaments (Fig. 3b and Extended Data Fig. 7b,c).

To validate whether Tpr indeed assembles earlier in the interphase assembly pathway, we used immunostaining of assembling NPCs with antibodies against Tpr and a later-assembling Nup, Nup62, and visualized them by 3D STED super-resolution microscopy. This single-NPC imaging demonstrated that a large number of assembling NPCs in the inner-core region in early G1 contain Tpr but not Nup62, whereas most of the assembled NPCs in the non-core region of the same cells contain both Tpr and Nup62 (Fig. 4a,c), confirming the surprisingly early recruitment of Tpr in the interphase assembly pathway. In addition to cells in early G1, we also examined cells later in interphase during S and G2 phase. Consistently, we found that a substantial number of NPCs also contain Tpr but not Nup62 (Fig. 4b,c), indicating that interphase assembly at later cell cycle stages follows the same order.

Fig. 4: Observation of single nuclear pores by super-resolution microscopy confirms the recruitment of Tpr precedes Nup62 in the interphase assembly pathway.


a
Mid/late telophase
Inner core
Nup62
Tpr
Merge
Non core
Nup62
Tpr
Merge
b
Interphase
Nup62
Tpr
Merge
c
Mid/late telophase
Inner core
Non core
Interphase
Frequency
1.0
0.8
0.6
0.4
0.2
0
****
Tpr+Nup62
Tpr only
Nup62 only
Tpr+Nup62
Tpr only
Nup62 only
Frequency
1.0
0.8
0.6
0.4
0.2
0
****
Tpr+Nup62
Tpr only
Nup62 only

a,**b**, Three-dimensional STED imaging of Nup62–mEGFP genome-edited cells stained with a GFP nanobody and an anti-Tpr antibody at mid/late telophase (**a**) and interphase (**b**). Scale bars: left, 10 μm; right, 1 μm. The regions of the nuclear envelope indicated by white lines (left) are flattened and displayed (right). Nuclear pores that contain only Tpr or Nup62 are indicated by red and green arrows, respectively. **c**, The frequency of nuclear pores that contain both Tpr and Nup62, only Tpr or only Nup62 at each cell cycle stage. Data are from 10 cells for each stage. Data are mean ± s.d. ****$P = 0.000017$ (mid/late telophase) and ****$P = 0.000080$ (interphase), unpaired two-tailed t-test.

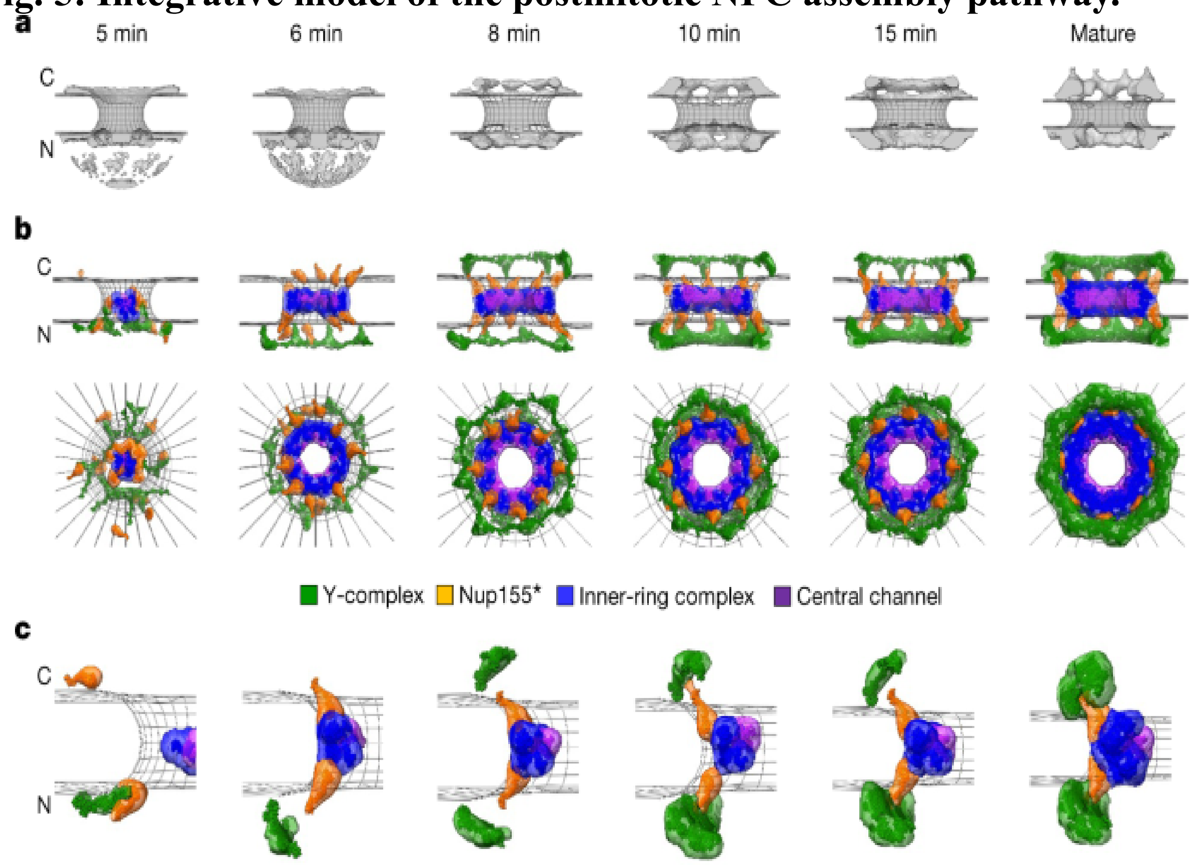
Source data

Integrative modelling of postmitotic assembly

To obtain a more comprehensive mechanistic view, we computed a spatiotemporal model of the macromolecular assembly pathway on the basis of our dynamic multimolecular stoichiometric data in combination with the available ultrastructural data about NPC assembly[11] and the partial pseudoatomic model of the mature human NPC[27,28]. We modelled only the postmitotic assembly, as it showed the kinetic hallmarks of a sequential process and is known to proceed by dilating an existing membrane pore with a smoothly growing proteinaceous density[11]. We focused on the Nups contained in the structural model of the human NPC, including Nup107 and Seh1 for the Y-complex as well as Nup93, Nup205 and Nup62 for the central ring or channel complex[27,28]. We constrained their copy numbers with our stoichiometry data for the time points for which correlative electron tomography data are available[11], enabling us to use the membrane shapes and associated protein densities from electron tomography to model the Nup configurations at these time points. To compute a spatiotemporal model of the assembly pathway, we generalized our integrative modelling method for determining static structures of macromolecular assemblies[29,30]. In outline (see Methods for details), we first computed ensembles of Nup configurations at discrete time points, connected pairs of these configurations at adjacent time points into assembly trajectories, and finally ranked the alternative trajectories by fit to our data.

We focused on the best fitting macromolecular assembly pathway that accounted for over 85% of the posterior model likelihood (the second-scoring model accounted for only 14.5%; Fig. 5 and Extended Data Fig. 8). This trajectory starts with the formation of a single nuclear ring, composed of eight Y-complexes, concomitantly with an initial accumulation of the FG-repeat protein Nup62 and the inner-ring complexes (the Nup93–Nup188–Nup155 and Nup205–Nup93–Nup155 complexes and Nup155) in the centre of the membrane hole (Fig. 5, 5 and 6 min). We experimentally validated that Nup155 and Nup188 assemble as the model predicted (Extended Data Figs. 9 and 10 and Supplementary Fig. 1). The cytoplasmic Y-complex is then added on the cytoplasmic side, before the second set of the Y-complex ring assembles on the nuclear side, again one eight-membered ring after another (Fig. 5, 8 and 10 min). In the centre of the pore, Nup62 dilates from an amorphous mass into a small ring and associates with inner-ring Nups to form the central ring complex. This 'nuclear-ring-first' assembly mechanism is consistent with the observation of an eight-fold symmetric protein density on the inner nuclear membrane at early stages of assembly[11].

Fig. 5: Integrative model of the postmitotic NPC assembly pathway.



a, Protein density (grey) overlaid with the nuclear envelope surface (wireframe model, grey) at each time point[12] used for integrative modelling. C, cytoplasm; N, nucleoplasm. The dome-like density in the nucleoplasmic side at 5 and 6 min is the noise from electron tomography. **b**, The best-scoring model of the postmitotic assembly pathway (top and side views). The uncertainty of each Nup localization is indicated by the density of the corresponding colour. The Y-complex is shown in green; an isolated fraction of Nup155 that is not forming a complex with Nup93, Nup188 and Nup205 is shown in orange; Nup93–Nup188–Nup155 and Nup205–Nup93–Nup155 complexes are in blue; and the Nup62–Nup58–Nup54 complex is in purple. **c**, The enlargement of one of the spokes. The model has a much higher score than all the other lower-scoring pathways (Extended Data Fig. 8). The pseudoatomic model of the native NPC structure[27,28] we imposed as the endpoint of the assembly pathway does not include the Y-complex-bound fraction of Nup205 or the Nup214-bound fraction of Nup62, and thus contains only 16 of the 40 copies of Nup205[33,34] and 32 of the 48 copies of Nup62[35] in the fully mature NPC. How the remaining 24 copies of Nup205 and 16 copies of Nup62 are assembled remains elusive.

Discussion

Our data revealed that the two NPC assembly pathways are markedly different. Although it had previously been shown that postmitotic and interphase assembly pathways have different molecular requirements[5,6,7,8,9,10,11,12], very little was known about the interphase pathway owing to its rare and sporadic nature. Although relative kinetics were available for some Nups[15,16], how key Nups composing cytoplasmic filaments, central channel, and nuclear basket, or multiple subunits of the Y-shaped and inner-ring complexes assemble in interphase had not been studied. Here we provide measurements on subunit composition and stoichiometry for both postmitotic and interphase assembly for ten Nups that represent all major building blocks of the NPC. Our data thus systematically reveal the main molecular differences between the two assembly pathways. During postmitotic assembly, the Y-complex is rapidly combined with components of the central ring, building the inner—and at this time already transport competent[13]—core of the pore before the addition of either

cytoplasmic or nuclear filament proteins. By contrast, during interphase assembly, the Y-complex is first combined with the nuclear filament protein Tpr and the base of the cytoplasmic filament Nup214, whereas the central ring complex is added much later.

The molecular pathway that we observed for interphase assembly makes new predictions about its unique, inside-out evaginating, mechanism. The cell first builds the two nuclear Y-rings and accumulates the material for their cytoplasmic counterparts and then combines them with the nuclear filament proteins, all preceding membrane fusion[12]. This mechanism suggests that the cytoplasmic Y-rings, including the base of the cytoplasmic filament Nup214, are already 'prebuilt' within the inner membrane evagination, where the small available volume would predict that they must be present in a very different structure than in the fully mature pore after fusion. Surprisingly, the central ring complex—which is a core structural element between the nuclear and cytoplasmic Y-rings in the mature pore, is added later—suggesting that during interphase assembly, nuclear transport control is added after the membrane fusion step. The unexpectedly early presence of Tpr suggests—although it does not prove—a potential role of Tpr in the initiation of interphase assembly. This long coiled-coil multimeric protein component could be involved in bending the inner nuclear membrane to create the invagination, potentially providing force by unfolding or restructuring a coiled-coil bundle[1,2]. Conversely, considering previous reports that Tpr together with the kinase ERK is required to control the spacing between NPCs[31], Tpr-based signalling might also be important in assembly site selection to prevent simultaneous assembly of multiple NPCs in close proximity with each other, which would probably result in abnormal distortions of the inner nuclear membrane.

Our data on interphase assembly in human cells furthermore revealed striking differences with budding yeast, which undergoes a closed mitosis and lacks the postmitotic pathway[32]: (1) the Tpr homologues Mlp1/2 assemble late in yeast; (2) human cells exhibit a synchronous assembly of the Y-complex components, whereas in yeast, the stem (for example, Nup107) and the head (for example, Seh1) of the Y-complex do not assemble simultaneously; and (3) the central ring components assemble after the Y-complex components in human cells, whereas they assemble together

in yeast[32]. These differences could be owing to the fundamental differences in structure (such as the presence of a nuclear lamina) and/or cell cycle remodelling of the nuclear envelope (closed versus open mitosis) between human and yeast cells.

Our dynamic stoichiometry data, combined with previous electron tomography data on membrane topology and protein volume, enabled us to predict the structures of the NPC assembly intermediates using integrative spatiotemporal modelling. We focused on the topologically simpler postmitotic assembly. The resulting model enables multiple new mechanistic predictions regarding postmitotic assembly that will help guide future work. For example, the model suggests that the central ring complex might be needed to prevent sealing of holes in the endoplasmic reticulum after mitosis, and that the hydrophobic FG-repeats of central channel Nups might have a role in the dilation of the small membrane hole into the larger, NPC-sized channel. Our pathway model was constrained by the fully assembled NPC structure[27,28] containing 16 copies of Nup205 and 32 copies of Nup62; thus, it does not address how additional copies of Nup205 and Nup62 that have been reported in recent structural studies and a proximity-dependent biotin identification study[27,33,34,35] may be assembled.

Our combined experimental and computational approach for determining macromolecular assembly pathways is likely to benefit from additional experimental data provided by emerging methods, such as 3D super-resolution microscopy for imaging the molecular architecture of the NPC intermediates[36,37] and dynamic super-resolution methods for mapping structural changes in real time[38]. We expect that this additional information will enable higher resolution modelling of even more complex assembly mechanisms, such as the interphase NPC assembly that involves much more substantial changes in membrane topology and protein conformation.

The need for better methods to determine dynamic and variable copy numbers of Nups reliably at the level of single protein complexes under physiological conditions is further highlighted by the fact that our live-cell-based approach resulted in copy number estimates that differ from estimates derived by other approaches for four out of the ten investigated Nups. The main methods used in the field, including quantitative mass spectrometry

from cell populations[23], fitting of single protein structures into highly averaged cryo-EM densities[33,34], as well as calibrated imaging of fluorescently tagged knock-in proteins in single living cells as used in our study, currently all have limitations in this regard. For example, cryo-EM based averaging often selects ‘complete’ NPCs, to obtain a ‘fully occupied’ average NPC model, whereas live-cell imaging data average NPC stoichiometry from all pores in one cell and thus include pore-to-pore variability, which can result in lower estimates. By contrast, whereas homozygous genome editing of a fluorescent tag ensures close to complete labelling and biological function of the fusion protein, it can still affect the expression level, which could lead to a different occupancy within the NPC.

Four out of the ten investigated Nups are relevant to discuss in the light of these methodological limitations—Nup153, Pom121, Nup205, and Nup358. For the first two, our estimates were much lower than expected from the structural model of the fully assembled state that we imposed as the end-state of our pathway model (Fig. 2 and Extended Data Figs. 1 and 2). We therefore regarded them as effectively expressed at subphysiological levels and normalized their stoichiometry to the expected number of copies in the mature pore for comparison with other Nups (Fig. 3a). For Nup205 and Nup358, our copy number estimates were 40–50% lower than expected, even though for Nup205 our cell lines express physiological levels of the fusion protein, whereas for Nup358 this correlates to a reduced level of expression (Extended Data Fig. 2). More work is needed to reliably measure the dynamic and variable composition of the NPC, especially during its assembly, and incorporate these data into modelling.

Beyond the assembly mechanism, the two pathways that we have mapped here begin to shed light on the role of the NPC during the evolution of the endomembrane system in eukaryotes[3,4]. We speculate that the modern NPC combines ancient membrane bending (for example, coiled-coil filaments, such as Tpr) and membrane hole plugging or dilation (for example, central ring complex and FG-repeat proteins) modules that were potentially previously used separately for extruding the cell surface or keeping transport channels open in the endomembranes around the genome. The key evolutionary innovation might lie in combining and controlling these activities, potentially by the eight-membered ring architecture of the nuclear

Y-complex. In future, comparing NPC assembly pathways in different species on the eukaryotic evolutionary tree might help us to understand how the assembly of complex modern protein machines reflects their evolutionary origins.

Methods

Cell culture

Wild-type HeLa Kyoto cells (RRID: CVCL_1922) were a gift from S. Narumiya; their genome was sequenced previously[39]. Cells were grown in high glucose Dulbecco's Modified Eagle's Medium (DMEM) containing 4.5 g l^{-1} d-glucose (Sigma Aldrich) supplemented with 10% fetal calf serum (FCS), 2 mM l-glutamine, 1 mM sodium pyruvate and 100 μg ml^{-1} penicillin and streptomycin at 37 °C and 5% CO_2. Cells tested negative for mycoplasma contamination by PCR every 2 or 3 months.

Genome editing

Monomeric enhanced GFP (mEGFP) and mCherry were inserted into the genome using zinc finger nucleases, CRISPR–Cas9 nickases, or Alt-R S.p. HiFi Cas9 Nuclease V3. The following six cell lines had been generated and published previously: Nup62–mEGFP[18], mEGFP–Nup107[12], mEGFP–Nup205[11], mEGFP–Nup214, mEGFP–Nup358 (also called RanBP2) and Tpr–mEGFP[17]. The following five cell lines were generated in this study: mEGFP–Seh1, Nup93–mEGFP, mEGFP–Nup153, Nup188–mEGFP and Pom121–mCherry. The guide RNA (gRNA) sequences used to generate these cell lines are summarized in Supplementary Table 2. For the Nup93–mEGFP cell line, CRISPR–Cas9 nickases and the donor plasmid were transfected by electroporation (Neon Transfection System, Thermo Fisher Scientific) instead of a polymer-mediated transfection reagent.

Western blot

Cells were lysed on ice in RIPA buffer (50 mM Tris-HCl, 150 mM NaCl, 1.0% Triton X-100, 1% sodium deoxycholate, 0.1% SDS, 2 mM EDTA, pH

7.5), supplemented with complete EDTA-free protease inhibitor cocktail (Sigma Aldrich), PhosSTOP (Sigma Aldrich), and 0.1 mM phenylmethylsulfonyl fluoride (Sigma Aldrich). The cell lysates were snap-frozen in liquid nitrogen and quickly thawed at 37 °C two times. The lysates were then centrifuged at 16,000*g* at 4 °C for 10 min and the supernatant was used for immunoblot analysis. Protein concentration was quantitated using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific). For Nup188 (Extended Data Fig. 10b), the cell lysates were run on NuPAGE 4–12% Bis-Tris Gels (Novex Life Technologies) and transferred onto PVDF membrane using the Bio-Rad transfer system. After blocking with 4% milk solution (nonfat milk powder in PBS + 0.05% Tween 20), the proteins were labelled with anti-Nup188 (A302–322A, Bethyl Laboratories, 1:5,000) and anti-γ-tubulin (T5192, Sigma Aldrich, 1:1,000) antibodies. Subsequently anti-rabbit IgG horseradish peroxidase (HRP)-conjugated secondary antibody (W4011, Promega) was used to detect the protein of interest with chemiluminescence reaction. For Nup205 and Nup358 (Extended Data Fig. 2), lysates were run on NuPAGE 3–8% Tris-acetate gels (Novex Life Technologies), and anti-Nup205 (ab157090, Abcam, 1:1,000), anti-Nup358 (HPA023960, The Human Protein Atlas, 1:1,000), and anti-vinculin (ab219649, Abcam, 1:1,500) antibodies were used to detect the proteins. Simple western assays were also performed in a Jess instrument (ProteinSimple) using 66–440 kDa capillary cartridges in accordance with the provider's instructions. The anti-Nup205, Nu358, and γ-tubulin antibodies were used at 1:50 dilution. For Nup205, mouse anti-γ-tubulin antibody (T5326, Sigma) was used instead of the rabbit anti-γ-tubulin antibody to adjust the signal intensity. The secondary antibodies used were the ones provided by the company at a ready-to-use dilution (goat anti-mouse HPR-conjugated secondary antibody, 040-655, Bio-Techne; goat anti-rabbit HPR-conjugated secondary antibody, 040-656, Bio-Techne).

FCS-calibrated live-cell imaging and estimation of Nup copy numbers per NPC

Live-cell imaging was performed using wild-type cells, wild-type cells transfected with mEGFP using Fugene6 (Promega), mEGFP–Nup107 genome-edited cells, or cells of another mEGFP–Nup genome-edited cell line. Cells were seeded in 8-well Lab-Tek Chambered Coverglass (Thermo

Fisher Scientific). On the day of live-cell imaging, DMEM was replaced with imaging medium: CO_2-independent medium without phenol red (Invitrogen) containing 20% FCS, 2 mM l-glutamine, and 100 μg ml^{-1} penicillin and streptomycin. The imaging medium was supplemented with 50 nM SiR (Hoechst) to stain DNA[24]. Cells were incubated inside the microscope-body-enclosing incubator at 37 °C for at least 30 min before imaging. For Nup188–mEGFP genome-edited cells, the following medium was also used instead of the imaging medium: 30 mM HEPES pH 7.4 containing 9.3 g l^{-1} minimum essential medium Eagle (Sigma Aldrich), 10% FCS, 1% MEM non-essential amino acids (Thermo Fisher Scientific, Gibco), and 100 μg ml^{-1} penicillin and streptomycin.

Calibrated imaging using FCS was carried out as described[22], using Fluctuation Analyzer 4G 150223 (https://www-ellenberg.embl.de/resources/data-analysis), FCSFitM v0.8 (https://git.embl.de/grp-ellenberg/FCSAnalyze), FCSCalibration v0.4.2 (https://git.embl.de/grp-ellenberg/FCSAnalyze), RStudio 1.1.383, R 3.4.1, and Python v3.6.8. In brief, the confocal volume was determined by performing FCS using a dye with known diffusion coefficient and concentration (Alexa Fluor 488 NHS ester; Thermo Fisher Scientific for mEGFP). To convert fluorescence intensity to the concentration, FCS was performed in cells that transiently express mEGFP alone, and a calibration curve was obtained by plotting the fluorescence intensity against concentration. The background fluorescence signal was measured in cells not expressing fluorescent proteins and subtracted.

To measure the concentration of Nups, mEGFP–Nup genome-edited cells in interphase were imaged in 3D using a confocal microscope (LSM780; Carl Zeiss, Oberkochen, Germany) and a 40× 1.2 NA C-Apochromat water immersion objective (Carl Zeiss) at 37 °C in a microscope-body-enclosing incubator, under the following conditions: 21 optical sections, section thickness of 2.0 μm, *z*-stacks every 1.0 μm, and *xy* pixel size of 0.25 μm. When the nuclear envelope is not perpendicular to the confocal plane of the 3D stacks, the fluorescence intensity at the nuclear envelope is non-isotropic in the point-spread function (PSF), which results in underestimation of the signal. To avoid such underestimation, a single plane was selected that contains the largest nuclear area in which the nuclear envelope is

perpendicular to the imaging plane and thus isotropic in the PSF. The fluorescence intensity of Nups was quantified on this single plane using the nuclear envelope mask with the width of three pixels (0.75 µm) that was generated from a SiR–DNA channel. Background fluorescence intensity was measured in wild-type cells without expressing any fluorescent proteins and subtracted. The Nup fluorescence intensity on the nuclear envelope was converted to the concentration using the calibration curve generated by FCS above. The number of Nups per μm^2 was calculated from the concentration and then divided by the NPC density per μm^2 measured by STED microscopy. This absolute quantification of Nup copy number with FCS calibration was done using 47 mEGFP–Nup107 genome-edited cells in interphase. For other mEGFP-Nups genome-edited cells, their Nup fluorescent intensities on the nuclear envelope were directly compared with the ones of mEGFP–Nup107 genome-edited cells on the same 8-well Lab-Tek Chambered Coverglass, and then their concentrations were determined using the intensity ratios to the mean intensity of mEGFP–Nup107 without using a FCS calibration curve. For Pom121–mCherry, the copy number was quantified independently by performing FCS using Alexa Fluor 568 NHS ester (Thermo Fisher Scientific) to measure the confocal volume and using the cells that transiently express mCherry alone to convert fluorescence intensity to the concentration.

Measurement of nuclear pore density by STED microscopy

Cells were fixed with 2.4% formaldehyde (Electron Microscopy Sciences) in PBS for 10 min, extracted with 0.4% Triton X-100 (Sigma Aldrich) in PBS for 5 min, and blocked with 5% normal goat serum (Life Technologies) in PBS for 10 min at room temperature. Subsequently, the cells were incubated overnight at 4 °C with a mouse anti-Nup62 (610497; BD Biosciences, 1.25 $\mu g\ ml^{-1}$) antibody, and then with an Abberior Star Red-conjugated anti-mouse IgG (Abberior, 0.5 $\mu g\ ml^{-1}$) for 30 min at room temperature. For Extended Data Fig. 4b, anti-GFP (Roche, 1:200) and anti-Elys (The Human Protein Atlas, 0.5 $\mu g\ ml^{-1}$) antibodies, Star Red-conjugated anti-rabbit IgG (Abberior, 0.5 $\mu g\ ml^{-1}$) and Star 580-conjugated anti-mouse IgG (Abberior, 0.5 $\mu g\ ml^{-1}$) were used. After multiple washes in PBS, cells were mounted in Vectashield (H-1500, Vector Laboratories). Super-resolution imaging was performed on a Leica SP8 3X STED microscope as described[12]. Images

were taken with a final optical pixel size of 20 nm, *z*-stacks of every 250 nm, and optical section thickness of 550 nm. Images were filtered with a Gaussian filter (kernel size: 0.5 × 0.5 pixels) for presentation purposes. The shrinkage of the nucleus caused by formaldehyde fixation and/or Vectashield mounting was quantified by comparing the volume of the nuclei of live cells with the ones of fixed cells. The shrinkage was 9.1 ± 2.6% (mean ± s.e.m., n = 36 cells). The NPC density was corrected for the nuclear shrinkage for the calculation of Nup copy number per NPC in Fig. 1d.

To quantify NPC density, the raw STED data were processed in ImageJ[40] with a mean filter (kernel size: 2 × 2 pixels) and a sliding paraboloid (radius: 5 pixels) for background subtraction. Detection of central peak positions for individual NPCs was carried out with the plugin TrackMate[41], using DoG detector and adjusting the detection threshold as the spot diameter size. The resulting 3D NPC coordinates were used to visualize and determine flat and curved regions of the nucleus. Using this map, circular and ellipsoidal regions of interest (ROIs) could then be selected in the flatter parts containing central NPC positions within the *z*-depth of approx. 500 nm, which corresponds to 2–3 microscopic slices in the images. The remaining signal outside the ROIs, as in curved regions or cytoplasmic structures were discarded from further analysis. NPC densities were calculated for each cell separately by dividing the number of NPCs within the selected ROIs by the corresponding ROI areas. For each cell line, the values were combined to calculate the mean and median NPC density values.

3D STED microscopy for visualizing single nuclear pore assembly intermediates

Cells were coated on 18 × 18 mm no. 1.5 square coverslips and synchronized by double thymidine arrest. After 10 h release from the second thymidine treatment, cells were fixed with 2.4% formaldehyde in PBS for 15 min, extracted with 0.25% Triton X-100 (Sigma Aldrich) in PBS for 15 min, and blocked with 2% bovine serum albumin (A2153; Sigma Aldrich) in PBS for 30 min at room temperature. Subsequently, the cells were incubated overnight at 4 °C with rabbit anti-Tpr (HPA019661; The Human Protein Atlas, 1:100) and a GFP nanobody (FluoTag-X4 anti-GFP nanobody Abberior Star 635P; N0304-Ab635P-L; NanoTag Biotechnologies, 1:50),

and then with an Alexa Fluor 594 goat anti-rabbit IgG (A-11037; Life Technologies, 1:250) for 30 min at room temperature. After multiple washes in PBS, cells were mounted in ProLong Gold Antifade Mountant (P10144; Invitrogen). Super-resolution imaging was performed on an Abberior STED/RESOLFT Expert Line microscope. Samples were imaged with an UPlan-S Apochromat 100× 1.4 NA oil-immersion objective on an IX83 stand (Olympus). Stimulated depletion was performed using a 775 nm pulsed laser (40 MHz) in combination with 594 and 640 nm pulsed excitation lasers in line switching mode. Fluorescence signal was detected using two separate Avalanche photo diodes with bandpass filters of 605–625 and 650–720 nm. The images were taken with a final optical pixel size of 35 nm, *z*-stacks of every 200 nm, and optical section thickness of 1,000 nm. For presentation purposes, images were filtered with PureDenoise plugin[42] in ImageJ, and lines with width of 175 nm were drawn along the nuclear envelope, flattened and shown in the figures.

Quantification of Nup copy number in the cytoplasm and the nucleoplasm and in non-core and core regions of the nuclear envelope

Mitotic cells were imaged and monitored from anaphase onset for 2 h in 3D by confocal microscopy. The microscopy setup and the imaging conditions are described above. Time-lapse imaging for mEGFP-tagged Nups was performed every 30 s. Photobleaching was negligible and thus not corrected. Time-lapse imaging for Pom121–mCherry was carried out every 60 s, and photobleaching was corrected by measuring a fluorescence signal decay in a neighbouring cell in the same field of view. Visualization of the chromosome surface in 3D was done in the Amira software package[43].

To measure Nup accumulation on the nuclear envelope, single planes were selected that contain the largest nuclear area at individual time points to avoid underestimation of the signal as mentioned earlier. The Nup intensity was quantified on the nuclear envelope mask with the width of 0.75 μm that was generated from a SiR–DNA signal at each time point. Except for Nup107 and Seh1, the Nup signal in the cytoplasm and nucleoplasm was measured and used as background. For Nup107 and Seh1, only the cytoplasmic signal was used as background because of their localization at

kinetochores. These background values were quantified at individual time points and subtracted from the Nup intensities on the nuclear envelope. The measured Nup intensity was converted into the concentration and then multiplied with nuclear surface area to calculate the total number of the Nups on the nuclear envelope. For Nup153 and Pom121, we did not convert the fluorescence intensity to the concentration as the cell lines were not fully validated to be homozygously tagged.

The Nup copy number was also calculated in the cytoplasm and the nucleoplasm during the first 2 h after anaphase onset. The cytoplasm mask was created by subtracting a mask of the nucleus generated from a SiR–DNA signal from a mask of the whole cell generated from a mEGFP–Nup signal. The mask for the nucleoplasm was created by eroding three pixels of the nuclear mask generated from a SiR–DNA signal. The Nup fluorescence intensity was quantified on these cytoplasmic and nucleoplasmic masks over time and then converted into the concentration. To calculate the total number of Nups in the cytoplasm and the nucleoplasm, the measured concentration was multiplied by the volume of the respective compartments. Cytosol volume at metaphase was quantified using the cytosolic signal of the Nups. The cytosolic GFP signal of some of the Nups (for example, Nup205, Nup214, Tpr and Nup358) were too low to precisely segment the cytosol in the *z*-slices close to the glass surface due to the relatively high background fluorescence signal. Therefore, we measured the cell volume using the brightest Nup62–mEGFP cell line (4,900 ± 330 μm^3 (mean ± s.d. from 8 cells)), and for the other Nup cell lines, we measured the cytosol area on a middle *z*-slice plane and calculated the ratios to the area for the Nup62 cell line (the ratios were 0.97, 0.97, 1.09, 1.14, 1.03, 1.07 and 1.02 for Nup107, Seh1, Nup205, Nup93, Nup214 and Nup358, respectively). From the measured volume of the Nup62 cell and the ratios of cytosol area to the Nup62 cell, we estimated the volume for the other cell lines (4,670, 4,660, 5,550, 6,030, 5,110, 5,420 and 5,080 μm^3 for Nup107, Seh1, Nup205, Nup93, Nup214 and Nup358 lines, respectively). For the cytoplasm volume after anaphase, we used the data that were measured previously in histone H2b–mCherry-expressing HeLa cells using fluorescently labelled dextran[17]. Assuming that the cell volume changes in the same degree during mitosis exit as the H2b–mCherry cell, we calculated the cytoplasmic volume of the Nup cell lines using the ratio to the volume at metaphase (5,300 μm^3 for the

H2b–mCherry cell[17]). The nucleoplasmic volume was quantified in each mEGFP–Nup knock-in cell line using a SiR–DNA signal at every time points as described previously[12].

Core regions were predicted on the nuclear envelope based on a previously described protocol using the core marker Lap-2α[12]. In brief, nuclear volume was segmented using SiR–DNA fluorescence signals that were processed with a 3D Gaussian filter and a multi-level thresholding. Nuclear volume was then divided into inner and outer volumes using the cutting plane that was constructed from the largest eigenvector and the second one orthogonal to the first vector of the pixel coordinates of the nuclear volume. Surface area of each nucleus was calculated and utilized to adjust the size of the inner and outer-core regions at individual time points. The previously defined criteria for being core and non-core regions[12] was applied. The position of inner and outer core was determined with respect to the intersection point of the largest eigenvector on the nuclear surface. The core region prediction was done in MATLAB (MathWorks).

Mathematical modelling for the nuclear pore assembly kinetics

Previous EM data showed that within 2 h after the onset of anaphase, postmitotic assembly is the dominant process in the non-core region, whereas slower interphase assembly predominates in the core region[12]. Assuming that this relation is also reflected in the live-cell Nup dynamics, we derived a mathematical model. We assumed that the observed total fluorescence intensity in the non-core, $n(t)$, and core region, $c(t)$, at time t after anaphase onset is a linear combination of the postmitotic, $\mathrm{pm}(t)$, and interphase assembly, $\mathrm{ip}(t)$, according to

$$n\left(t\right)={f}_{n}{\rm{pm}}\left(t\right)+\left(1-{f}_{n}\right){\rm{ip}}\left(t\right)$$

(1)

$$c\left(t\right)={f}_{c}{\rm{pm}}\left(t\right)+\left(1-{f}_{c}\right){\rm{ip}}\left(t\right).$$

(2)

The fraction of postmitotic assembly in the non-core and core regions are denoted f_n and f_c, respectively. To test this assumption and obtain an estimate of the fractions, we used a model that accounts for the observed sigmoid-like kinetics

$${\rm{pm}}(t)=\frac{{(t-{d}_{n/c})}^{{n}_{p}}}{{(t-{d}_{n/c})}^{{n}_{p}}+{K}_{p}^{{n}_{p}}},\,{\rm{for}}\,\,t\ge {d}_{n/c}$$
(3)
$${\rm{ip}}(t)=\frac{{\left(t-{d}_{n/c}\right)}^{{n}_{i}}}{{\left(t-{d}_{n/c}\right)}^{{n}_{i}}+{K}_{i}^{{n}_{i}}},\,{\rm{for}}\,\,t\ge {d}_{n/c}$$
(4)

and ${\rm{pm}}\left(t < {d}_{n/c}\right)={\rm{ip}}\left(t < {d}_{n/c}\right)=0.$ The parameters n_p, K_p and n_i, K_i characterize the postmitotic and interphase kinetics, respectively. The parameter ${d}_{n/c}$ accounts for an additional delay in NPC initiation. In the non-core region we assumed $d_n = 0$; in the core region due to the presence of kinetochore microtubule fibres[26], $d_c > 0$. The model is used to derive the underlying postmitotic and interphase assembly kinetics. Using equations (1 and 2), we obtain

$${\rm{ip}}\left(t\right)=\frac{{f}_{n}c\left(t\right)-{f}_{c}n\left(t\right)}{{f}_{n}-{f}_{c}}.$$
(5)
$${\rm{pm}}\left(t\right)=\frac{\left(1-{f}_{c}\right)n\left(t\right)-\left(1-{f}_{n}\right)c\left(t\right)}{{f}_{n}-{f}_{c}}.$$
(6)

Parameter estimation

For each Nup, there are two parameters that define the interphase (n_i, K_i) and postomitic (n_p, K_p) assembly, respectively. The fractions f_n and f_c and the delay d_c in the core region were estimated globally for all Nups. In total, we

have 43 parameter, $4 \times 10 = 40$ parameters describing the assembly kinetics and 3 global parameters, fitted to 4,446 data points.

In detail, to find the model parameters we minimized the mean squared distance between data and model for all the time points M

$${\chi }^{2}=\mathop{\sum }\limits_{j=1}^{M}\left({\left(\frac{N\left({t}_{j}\right)-n\left({t}_{j}\right)}{{\sigma }_{N}\left({t}_{j}\right)}\right)}^{2}+{\left(\frac{C\left({t}_{j}\right)-c\left({t}_{j}\right)}{{\sigma }_{C}\left({t}_{j}\right)}\right)}^{2}\right).$$
(7)

$N\left({t}_{j}\right)$ and $C\left({t}_{j}\right)$ are the mean background-subtracted and normalized fluorescence intensities in the non-core and core region with standard deviation ${\sigma }_{N}\left({t}_{j}\right)$ and ${\sigma }_{C}\left({t}_{j}\right)$, respectively, at time point $t = t_j$. We subtracted a background computed from the average of the first 3 time points. The data were normalized with the average value between 100 and 120 min after anaphase. In a first step, for each protein and cell line, we estimated the postmitotic fractions in the core and non-core region and the kinetic parameters. Overall, we computed 61 parameters (6 parameters per protein plus one delay parameter). For the postmitotic fractions, we obtained on average $f_n = 0.857$ [0.76, 0.95] and $f_c = 0.295$ [0.17, 0.4], where the number in brackets indicates the 95% confidence interval estimated using the profile likelihood method[44]. Importantly, the obtained postmitotic fractions are well in agreement with the previously reported estimates obtained from EM data[12]. The delay in pore formation between core and non-core region was estimated by systematically varying d_c from 0 to 6 min in steps of 1 min. A value of $d_c =$ 2 min, gave optimal result. In a second step, we used the previously estimated average postmitotic fractions, f_n and f_c, and d_c and recomputed the kinetics parameters for each protein. The model with reduced parameters well agrees with the data (Extended Data Figs. 5 and 6, $R^2 > 0.99$). To verify if the choice of common postmitotic fractions for all Nups is valid, we computed the Baysian information criterion (BIC)[45]. The difference in BIC between the model with reduced parameters, 43 parameters, compared to the

full model, 61 parameters, was −7, indicating that the model with reduced parameters is justified. The obtained parameter values are listed in Supplementary Table 3.

Median assembly time and duration of assembly

One can think of an assembly curve as a cumulative distribution function and its derivative as the corresponding probability density function (PDF) for binding events. The median assembly time, i.e. the time where 50% of binding events have occurred, is K_i and K_p, for the interphase and postmitotic assembly respectively. We consider the time intrinsic to the assembly mechanism and independent of the initiation delay d_c. We further define the duration of assembly as the time interval where 80% of binding events occur, i.e. from a fraction of $\alpha_1 = 0.1$ up to $\alpha_2 = 0.9$, one obtains (see also Extended Data Fig. 7a)

$$\Delta {T}_{p}={K}_{p}\left({\left(\frac{{\alpha }_{2}}{1-{\alpha }_{2}}\right)}^{\frac{1}{{n}_{p}}}-{\left(\frac{{\alpha }_{1}}{1-{\alpha }_{1}}\right)}^{\frac{1}{{n}_{p}}}\right)$$
(8)
$$\Delta {T}_{i}={K}_{i}\left({\left(\frac{{\alpha }_{2}}{1-{\alpha }_{2}}\right)}^{\frac{1}{{n}_{i}}}-{\left(\frac{{\alpha }_{1}}{1-{\alpha }_{1}}\right)}^{\frac{1}{{n}_{i}}}\right).$$
(9)

The duration of assembly quantifies the width of the binding events PDF.

In Extended Data Fig. 7c we see a strong positive correlation between median assembly time and duration of assembly for the postmitotic assembly. This suggests a sequential assembly mechanism. The rationale is that for a strict sequential pathway, the binding events PDF for subsequent Nup is a convolution of all previous binding events PDF and so will broaden for later binding proteins. For example, if an early protein has an assembly duration of 1 h, within this time window binding sites for the subsequent protein will continue to appear. Therefore, the subsequent protein in the sequence will also accumulate for at least 1 h.

For the simplified case of irreversible sequential assembly and linear rate constants, a positive correlation can be demonstrated. This correlation is independent on whether the initiation is synchronous or spread within a time-period. For simplicity reasons, we omit the fact that each Nup binds in multiple copies to the NPC. A Nup P_i binds with rate constant κ_i to a NPC intermediate X_{i-1} according to the reaction scheme in Supplementary Fig. 2.

The corresponding system of ordinary differential equations is

$$\frac{d{X}_{0}}{{dt}}=g(t)-{k}_{1}{X}_{0}$$
$$\frac{{\rm{d}}{X}_{i}}{{\rm{d}}t}={k}_{i}{X}_{i-1}-{k}_{i+1}{X}_{i},\text{for}n > i > 0$$
$$\frac{{\rm{d}}{X}_{n}}{{\rm{d}}t}={k}_{n}{X}_{n-1}$$

with $X_i(0) = 0$, $ {k}_{i}={\kappa }_{i}{P}_{i}^{\text{free}}$, and an excess of free Nup ${P}_{i}^{\text{free}}\approx {\rm{constant}}$. The function g accounts for the appearance of NPC initiation sites after anaphase. We assume that NPC initiation is completed within a finite time after anaphase and neglect the slowly and continuous appearance of pores in G1—that is, ${\int }_{0}^{\infty }g\left(t\right){\rm{d}}t={\rm{constant}}$. The amount of ith Nup bound to a complex is ${P}_{i}=\mathop{\sum }\limits_{j=i}^{n}{X}_{i}$ and its time derivative ${f}_{i}\left(t\right)=\frac{{\rm{d}}{P}_{i}}{{\rm{d}}t}={k}_{i}{X}_{i-1}$ is proportional to the binding events PDF. We can quantify the assembly time τ_i and assembly duration θ_i by the mean and standard deviation of the binding events PDF, respectively[46],

$${\tau }_{i}=\frac{{\int }_{0}^{{\rm{\infty }}}t{f}_{i}(t){\rm{d}}t}{{\int }_{0}^{{\rm{\infty }}}{f}_{i}\left(t\right){\rm{d}}t}$$

and

$${\theta }_{i}=\sqrt{\frac{{\int }_{0}^{{\rm{\infty }}}{t}^{2}{f}_{i}(t){\rm{d}}t}{{\int }_{0}^{{\rm{\infty }}}{f}_{i}\left(t\right){\rm{d}}t}-{\tau }_{i}^{2}}$$

After integration by parts, one obtains

$${\tau }_{i}={\tau }_{g}+\mathop{\sum }\limits_{j=1}^{i}\frac{1}{{k}_{j}}$$
(10)

and

$${\theta }_{i}=\sqrt{{{\theta }_{g}}^{2}+\mathop{\sum }\limits_{j=1}^{i}\frac{1}{{k}_{j}^{2}}}.$$
(11)

Where $({\tau }_{g}=\frac{{\int }_{0}^{\infty }tg(t){\rm{d}}t}{{\int }_{0}^{\infty }g(t){\rm{d}}t})$ and $({\theta }_{g}=\sqrt{\frac{{\int }_{0}^{\infty }{t}^{2}g(t){\rm{d}}t}{{\int }_{0}^{\infty }g(t){\rm{d}}t}-{\tau }_{g}^{2}})$ are the mean and standard deviation of the initiation function, respectively. From equations (10 and 11), it is clear that $({\tau }_{i+1} > {\tau }_{i})$ and $({\theta }_{i+1} > {\theta }_{i})$ for any parameter combinations and independently on how synchronous the initiation of pores is. This shows that for a strictly sequential pathway a linear correlation between assembly time and duration is expected.

Integrative modelling of the NPC assembly pathway

A model of the assembly pathway is defined by a series of static structures, including a static structure at each sampled time point along the assembly process. Therefore, we model the NPC assembly by first modelling static structures at each time point, independently from each other. We then enumerate alternative assembly pathways and rank them based on the static structure scores and plausibility of transitions between successive static structures.

Integrative modelling of static structures at each time point

The static structures are modelled by standard integrative structure modelling[30] as follows.

Representing a static structure model

The time points correspond to times with available electron tomography protein densities[11]: 5 min, 6 min, 8 min, 10 min and 15 min after anaphase onset. We divide the mature NPC structure (PDB: 5A9Q and 5IJO) into eight spokes and further divide each spoke into a set of rigid subcomplexes, including the Y-complex, the inner-ring Nup205–Nup155–Nup93 subcomplex, the inner-ring Nup93–Nup188–Nup155 subcomplex, and the central channel Nup62–Nup58–Nup54 subcomplex. For each domain, we coarse-grained the structure by grouping ten consecutive amino acid residues into a single bead at the centre of mass of those residues. Each subcomplex is represented as a rigid body. The nuclear envelope is represented as a fixed toroid surface embedded in two parallel planes. Thus, the variables of the model include the Euclidean coordinates of the Nup subcomplexes and the copy number of each Nup subcomplex.

We set the inner pore diameter and minor radius of the pore at each time point to the mean of previously determined nuclear envelope cross sections[11] with a pore diameter of 51.5 nm, 58.4 nm, 72.7 nm, 84.6 nm, 79.8 nm and 87 nm; and minor radius of 21.4 nm, 21.2 nm, 21.5, 20.3 nm, 17.1 nm and 15 nm for time points at 5 min, 6 min, 8 min, 10 min, 15 min and the mature pore, respectively.

Scoring a static structure model

The copy numbers of the NPC subcomplexes at each time point were restrained by a Gaussian function with mean and variance determined by the single-cell traces presented in this study. The relative likelihood of a set of copy numbers is proportional to the product of individual Gaussian likelihoods.

Distances between pairs of Nups that are in contact with each other in the native NPC structure[27,28] were restrained by a harmonic Gō-like model[47]. Inter-subcomplex contacts within 5 nm in the mature structure were restrained by a harmonic function (strength 0.01 kcal mol^{-1} Å). Each Gō-like scoring term was scaled at each time point, from zero at the first time point to full strength at the mature pore time point. Distances between all pairs of Nups were also restrained by a harmonic excluded volume restraint (strength 0.01 kcal mol^{-1} $Å^{-1}$). Proximity between Nup domains containing a

membrane interacting ALPS-motif and the nuclear envelope was restrained by a harmonic term (strength 0.1 kcal mol^{-1} $Å^{-1}$), based on their sequences. Overlap between the Nups and nuclear envelope surface was avoided by imposing a harmonic repulsion between the Nups and nuclear envelope surface (strength of 0.01 kcal mol^{-1} $Å^{-1}$).

The shape of a static structure was restrained by a correlation coefficient between the model and electron tomography protein density[11]. The forward model density was represented by fitting each Nup subcomplex with a Gaussian mixture model of two components per subcomplex copy using the gmconvert utility[48]. Similarly, the electron tomography protein densities at each time point were represented with a Gaussian mixture model with 150 components fit to the experimental density.

Sampling static structure models

A state of the NPC at any given time point is defined by the copy numbers and coordinates of its components. Only copy number assignments and structures consistent with the C-8 symmetries were sampled. The assumption that the subcomplexes preform with 8-fold multiplicity before assembling into the NPC is supported by our previous electron microscopy analysis[11]. The averaged electron tomograms of the intermediates demonstrated 8-fold symmetry of the outer-ring complex at 5, 6, 8, 10, and 15 min after anaphase onset and 8-fold symmetry of the inner-ring complex at 10 and 15 min after anaphase onset[11], indicating that the majority of intermediates have approximate 8-fold symmetry for the outer- and inner rings. However, there are likely to be variations in the copy numbers at the single pore level. We only sampled structures for the top 20-scoring Nup copy number combinations. Each sampling started with the mature pore structure, followed by applying 10^6 Monte Carlo moves. These moves included rotational and translational perturbations to each Nup subcomplex, drawn from a uniform distribution in the range from −0.04 to +0.04 radians and from −4 to +4 Å, respectively.

Modelling the assembly pathway

With the static structure models in hand, we connect them into complete alternative assembly pathways, as follows.

Each pathway is represented by a static structure at each sampled time point, starting with $t = 5$ min and culminating in the native structure; we do not model the completely disassembled NPC. The score of a pathway is the sum of the scores for the static structures on the pathway (defined above) and transitions between them. A transition score is uniform for all allowed transitions. A transition between two successive static structures is allowed if the subcomplexes in the first structure are included in the second structure. All possible pathways were enumerated, scored, and ranked. The best-scoring pathways were extracted for further analysis (Fig. 5). Molecular visualization was performed with UCSF ChimeraX[49].

Immunofluorescence for visualizing Nup155 during assembly

Cells were treated, fixed and blocked as described in the previous section (STED microscopy for visualizing single nuclear pore assembly intermediates). Afterwards, cells were incubated overnight at 4 °C with rabbit anti-Nup155 (HPA037775; The Human Protein Atlas, 1:100), and then with an Abberior Star 635P goat anti-rabbit IgG (ST635P-1002-500UG, Abberior GmbH, 1:250) for 30 min at room temperature. After multiple washes in PBS, cells were mounted in ProLong Gold Antifade Mountant with DAPI (P36941, Invitrogen). Imaging was performed in 3D using a confocal microscope (LSM780; Carl Zeiss) and a 63× 1.4 NA Plan-Apochromat objective (Carl Zeiss) under the following conditions: 31 optical sections, section thickness of 0.7 µm, *z*-stacks of every 0.39 µm, and *xy* pixel size of 0.13 µm.

Sample size determination and statistical analysis

For quantitative imaging in Fig. 1a,d, the data were from 4, 4, 4, 2, 3, 2, 3 and 2 independent experiments for Nup107, Seh1, Nup205, Nup93, Nup62, Nup214, Tpr and Nup358, respectively. STED imaging in Fig. 4a and Extended Data Fig. 4a and live imaging in Extended Data Fig. 10 were from two independent experiments. For dynamic quantitative imaging in Fig. 3, the data were from 4, 4, 4, 2, 3, 4, 3, 2, 2 and 4 independent experiments for

Nup107, Seh1, Nup205, Nup93, Nup62, Nup214, Tpr, Nup358, Nup153 and Pom121, respectively. STED imaging in Figs. 1b,c and 4b, Southern blotting in Extended Data Fig. 1 and immunofluorescence microscopy in Extended Data Fig. 9 are from single experiments. Statistical analyses were performed only after all the data were taken. Sample sizes for each experiment are indicated in figure legends. Sample sizes were based on pilot experiments to determine the number of cells required to observe stable population averages with high Pearson's correlation between replicates. No blinding and randomization was done, as this study does not involve animals or human participants. Samples were organized into groups based on cell lines. Cells were imaged in randomly chosen fields of view per experiment. All imaged cells were further analyzed. Videos of dividing cells with rotating nuclei are removed from the analysis, because we cannot properly assign the non-core and core regions.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Fluorescence images were deposited to the Image Data Resource (IDR; https://idr.openmicroscopy.org/) under accession number idr0115. Our integrative spatiotemporal model of the postmitotic assembly of the human NPC has been deposited to the nascent database of integrative structures PDB-Dev (https://pdb-dev.wwpdb.org/) under accession number PDBDEV_00000142. Source data are provided with this paper.

Code availability

All source code is accessible on Github repository (for quantifying Nup intensity in core and non-core regions: https://github.com/mjh1m22/Quantitative_map_nups_Otsuka_Nature_2022; for decomposing postmitotic and interphase assembly kinetics: https://github.com/manerotoni/npc_assembly_Otsuka_2022). Integrative

Modeling Platform (IMP) is an open-source program freely available under the LGPL license at http://integrativemodeling.org; all input files, scripts, and output files are available at http://integrativemodeling.org/npcassembly.

References

1. Hampoelz, B., Andres-Pons, A., Kastritis, P. & Beck, M. Structure and assembly of the nuclear pore complex. *Annu. Rev. Biophys.* **48**, 515–536 (2019).

2. Lin, D. H. & Hoelz, A. The structure of the nuclear pore complex (an update). *Annu. Rev. Biochem.* **88**, 725–783 (2019).

3. Baum, D. A. & Baum, B. An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76 (2014).

4. Otsuka, S. & Ellenberg, J. Mechanisms of nuclear pore complex assembly - two different ways of building one molecular machine. *FEBS Lett.* **592**, 475–488 (2018).

5. Franz, C. et al. MEL-28/ELYS is required for the recruitment of nucleoporins to chromatin and postmitotic nuclear pore complex assembly. *EMBO Rep.* **8**, 165–172 (2007).

6. Doucet, C. M., Talamas, J. A. & Hetzer, M. W. Cell cycle-dependent differences in nuclear pore complex assembly in metazoa. *Cell* **141**, 1030–1041 (2010).

7. Talamas, J. A. & Hetzer, M. W. POM121 and Sun1 play a role in early steps of interphase NPC assembly. *J. Cell Biol.* **194**, 27–37 (2011).

8. Funakoshi, T., Clever, M., Watanabe, A. & Imamoto, N. Localization of Pom121 to the inner nuclear membrane is required for an early step of interphase nuclear pore complex assembly. *Mol. Biol. Cell* **22**, 1058–1069 (2011).

9. Vollmer, B. et al. Nup153 recruits the Nup107–160 complex to the inner nuclear membrane for interphasic nuclear pore complex

assembly. *Dev. Cell* **33**, 717–728 (2015).

10. Rampello, A. J. et al. Torsin ATPase deficiency leads to defects in nuclear pore biogenesis and sequestration of MLF2. *J. Cell Biol.* **219**, e201910185 (2020).

11. Otsuka, S. et al. Postmitotic nuclear pore assembly proceeds by radial dilation of small membrane openings. *Nat. Struct. Mol. Biol.* **25**, 21–28 (2018).

12. Otsuka, S. et al. Nuclear pore assembly proceeds by an inside-out extrusion of the nuclear envelope. *eLife* **5**, e19071 (2016).

13. Dultz, E. et al. Systematic kinetic analysis of mitotic dis- and reassembly of the nuclear pore in living cells. *J. Cell Biol.* **180**, 857–865 (2008).

14. Otsuka, S., Szymborska, A. & Ellenberg, J. Imaging the assembly, structure, and function of the nuclear pore inside cells. *Methods Cell. Biol.* **122**, 219–238 (2014).

15. D'Angelo, M. A., Anderson, D. J., Richard, E. & Hetzer, M. W. Nuclear pores form de novo from both sides of the nuclear envelope. *Science* **312**, 440–443 (2006).

16. Dultz, E. & Ellenberg, J. Live imaging of single nuclear pores reveals unique assembly kinetics and mechanism in interphase. *J. Cell Biol.* **191**, 15–22 (2010).

17. Cai, Y. et al. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature* **561**, 411–415 (2018).

18. Farrants, H. et al. Chemogenetic control of nanobodies. *Nat. Methods* **17**, 279–282 (2020).

19. Koch, B. et al. Generation and validation of homozygous fluorescent knock-in cells using CRISPR–Cas9 genome editing. *Nat. Protoc.* **13**, 1465–1487 (2018).

20. Guglielmi, V., Sakuma, S. & D'Angelo, M. A. Nuclear pore complexes in development and tissue homeostasis. *Development* **147**, dev183442 (2020).

21. Wachsmuth, M. et al. High-throughput fluorescence correlation spectroscopy enables analysis of proteome dynamics in living cells. *Nat. Biotechnol.* **33**, 384–389 (2015).

22. Politi, A. Z. et al. Quantitative mapping of fluorescently tagged cellular proteins using FCS-calibrated four-dimensional imaging. *Nat. Protoc.* **13**, 1445–1464 (2018).

23. Ori, A. et al. Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol. Syst. Biol.* **9**, 648 (2013).

24. Lukinavicius, G. et al. SiR-Hoechst is a far-red DNA stain for live-cell nanoscopy. *Nat. Commun.* **6**, 8497 (2015).

25. Maeshima, K. et al. Cell-cycle-dependent dynamics of nuclear pores: pore-free islands and lamins. *J. Cell Sci.* **119**, 4442–4451 (2006).

26. Vietri, M. et al. Spastin and ESCRT-III coordinate mitotic spindle disassembly and nuclear envelope sealing. *Nature* **522**, 231–235 (2015).

27. von Appen, A. et al. In situ structural analysis of the human nuclear pore complex. *Nature* **526**, 140–143 (2015).

28. Kosinski, J. et al. Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* **352**, 363–365 (2016).

29. Kim, S. J. et al. Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).

30. Rout, M. P. & Sali, A. Principles for integrative structural biology studies. *Cell* **177**, 1384–1403 (2019).

31. McCloskey, A., Ibarra, A. & Hetzer, M. W. Tpr regulates the total number of nuclear pore complexes per cell nucleus. *Genes Dev.* **32**, 1321–1331 (2018).

32. Onischenko, E. et al. Maturation kinetics of a multiprotein complex revealed by metabolic labeling. *Cell* **183**, 1785–1800 (2020).

33. Mosalaganti, S. et al. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **376**, eabm9506 (2022).

34. Petrovic, S. et al. Architecture of the linker-scaffold in the nuclear pore. *Science* **376**, eabm9798 (2022).

35. Kim, D. I. et al. Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc. Natl Acad. Sci. USA* **111**, E2453–E2461 (2014).

36. Thevathasan, J. V. et al. Nuclear pores as versatile reference standards for quantitative superresolution microscopy. *Nat. Methods* **16**, 1045–1053 (2019).

37. Sabinina, V. J. et al. Three-dimensional superresolution fluorescence microscopy maps the variable molecular architecture of the nuclear pore complex. *Mol. Biol. Cell* **32**, 1523–1533 (2021).

38. Gwosch, K. C. et al. MINFLUX nanoscopy delivers 3D multicolor nanometer resolution in cells. *Nat. Methods* **17**, 217–224 (2020).

39. Landry, J. J. et al. The genomic and transcriptomic landscape of a HeLa cell line. *G3* **3**, 1213–1224 (2013).

40. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

41. Tinevez, J. Y. et al. TrackMate: an open and extensible platform for single-particle tracking. *Methods* **115**, 80–90 (2017).

42. Luisier, F., Vonesch, C., Blu, T. & Unser, M. Fast interscale wavelet denoising of Poisson-corrupted images. *Signal Process.* **90**, 415–427 (2010).

43. Pruggnaller, S., Mayr, M. & Frangakis, A. S. A visualization and segmentation toolbox for electron microscopy. *J. Struct. Biol.* **164**, 161–165 (2008).

44. Raue, A. et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929 (2009).

45. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).

46. Heinrich, R., Neel, B. G. & Rapoport, T. A. Mathematical models of protein kinase signal transduction. *Mol. Cell* **9**, 957–970 (2002).

47. Go, N. & Abe, H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers* **20**, 991–1011 (1981).

48. Kawabata, T. Gaussian-input Gaussian mixture model for representing density maps and atomic models. *J. Struct. Biol.* **203**, 1–16 (2018).

49. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

Acknowledgements

We thank the EMBL Advanced Light Microscopy Facility (ALMF) for their support in STED super-resolution microscopy; S. Schnorrenberg for help with 3D STED; A. Brunner for help with automated live-cell imaging; B. Webb for help with IMP and J. Baker-Lepain for managing the Wynton computer cluster at QBI@UCSF. This work was supported by grants from the Baden Wuerttemberg Foundation (J.E. and A.S.), NIH/NIGMS R01GM083960 (A.S.), NIH/NIGMS P41GM109824 (A.S.), NIH/NIGMS R01GM112108 (A.S.) and the European Molecular Biology Laboratory

(EMBL; S.O., A.Z.P., A.R., M.J.H., M.K., A.C., B.K. and J.E.). S.O. and A.R. were further supported by the EMBL Interdisciplinary Postdoc Programme (EIPOD) under Marie Curie Actions COFUND. S.O. was additionally supported by a Japan Society for the Promotion of Science fellowship (postdoctoral fellowship for research abroad).

Author information

Author notes

1. Shotaro Otsuka

   Present address: Max Perutz Labs, University of Vienna and the Medical University of Vienna, Vienna Biocenter (VBC), Vienna, Austria

2. Antonio Z. Politi

   Present address: Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany

3. M. Julius Hossain

   Present address: Centre for Cancer Immunology, Faculty of Medicine, University of Southampton, Southampton, UK

4. Birgit Koch

   Present address: Max Planck Institute for Medical Research, Heidelberg, Germany

Authors and Affiliations

1. Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany

   Shotaro Otsuka, Wanlu Zhang, Antonio Z. Politi, Arina Rybina, M. Julius Hossain, Moritz Kueblbeck, Andrea Callegari, Birgit

Koch, Natalia Rosalia Morero & Jan Ellenberg

2. Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA, USA

   Jeremy O. B. Tempkin & Andrej Sali

Contributions

S.O., J.O.B.T., A.S. and J.E. designed the project. S.O. performed all the fluorescence microscopy experiments and analyses, except for the experiments in Fig. 4 and Extended Data Figs. 9 and 10. J.O.B.T. performed integrative structural modelling. W.Z. performed 3D STED, immunofluorescence time-course microscopy and live imaging shown in Fig. 4 and Extended Data Figs. 9 and 10. A.Z.P. carried out mathematical modelling for the nuclear pore assembly kinetics. A.R. developed an analysis pipeline for nuclear pore density measurement. M.J.H. established a computational image analysis pipeline to quantify fluorescence intensities in non-core and core regions in 3D time-lapse images. W.Z., M.K., A.C., B.K. and N.R.M. generated and validated genome-edited cell lines. A.S. and J.E. supervised the work. S.O., J.O.B.T., A.S. and J.E. wrote the paper. All authors contributed to the analysis and interpretation of data and provided input on the manuscript.

Corresponding authors

Correspondence to Shotaro Otsuka or Jan Ellenberg.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Valérie Doye, Paul Maddox and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Southern blotting of genome-edited cell clones.

Genomic DNA of each cell clone was digested with restriction enzymes and the fragments were detected by probes for Nup93 (**a**), Seh1 (**b**), Nup153 (**c**), and Pom121 (**d**), as well as GFP (**a**–**c**) and mCherry (**d**). The size of the fragments and the probe-binding regions are illustrated in the bottom panels. The clones indicated in red were used for this study. We regarded the clone #32 of Pom121-mCherry as subhomozygous because the protein abundance was much less than expected although the blot indicated homozygous tagging. The following Nup cell lines were validated to be homozygously-tagged in previous reports (Tpr, Nup214 and Nup358[17]; Nup107[12]; Nup205[11]; Nup62[18]).

Extended Data Fig. 2 Immunoblot analysis of mEGFP-Nup205 (a,b) and mEGFP-Nup358 (c,d) cell lysates.

The membranes were cut at the position of around 200 kDa. The membranes above 200 kDa were blotted with antibodies against Nup205 (**a**) and Nup358 (**c**), and the membranes below 200 kDa were blotted with an antibody against Vinculin for loading control. The clone 081 (**a**) and 097 (**c**) are the cell lines that are used in this study. (**b**,**d**) The relative intensity of the bands

for GFP-tagged Nups over endogenous Nups. The plots for Nup205 and Nup358 are from 9 and 8 pairs of bands (GFP-tagged vs endogenous) from 4 and 3 independent experiments, respectively. The mean is depicted as a horizontal line and the whiskers show the standard deviation.

Source data

Extended Data Fig. 3 NPC assembly relies on and consumes almost half of the material inherited from the mother cell within one hour after mitosis.

FCS-calibrated 3D confocal microscopy was performed as in Fig. 2, and the Nup copy number was quantified in whole dividing cells (for details see Methods). The number of Nups in cytoplasm (dark blue), nucleoplasm (medium dark blue) and nuclear envelope (light blue) are plotted against time after anaphase onset. The plot is the mean of 15, 20, 13, 14, 22, 22, 19, and 24 cells for Nup107, Seh1, Nup205, Nup93, Nup62, Nup214, Tpr and Nup358, respectively.

Source data

Extended Data Fig. 4 GFP-tagged Nups are recruited to the NPCs rather than nonspecifically accumulated on the NE.

a, Three-dimensional stimulated emission depletion (3D-STED) imaging of a Nup62-mEGFP genome-edited cell at early telophase. The region of the nuclear envelope indicated by a white line in the top image is flattened and shown in the bottom. Scale bars, 1 µm. The density of the Nup62 spots was 14.6 ± 2.6 NPCs/μm^2 (from 6 cells), which is consistent with our previous EM observation of the NPC density of 12–16 NPCs/μm^2 in early telophase cells[11]. **b**, 2D-STED imaging of a mEGFP-Nup107 genome-edited cell at early telophase that were stained with anti-GFP and anti-Elys antibodies. Enlarged images indicated by white boxes in the top images are shown in the bottom panel. Scale bars, 1 µm.

Source data

Extended Data Fig. 5 Postmitotic and interphase assembly are spatially distinguished for the first hour after mitotic exit and thus their kinetics can be decomposed.

a, Time-lapse 3D imaging of Nup93-mEGFP genome-edited cell. Single confocal sections of SiR-DNA and GFP channels are shown. Images were filtered with a median filter (kernel size: 0.25 × 0.25 µm). Scale bar, 10 µm. Time after AO is indicated. **b**, The fluorescence intensities at non-core (brown) and inner-core (yellow) regions were quantified. Dots represent the average and s.d. of measurements from 14 cells. The intensities were fitted with a sequential model of NPC assembly (bold lines) that allows for different population and rate constants for postmitotic and interphase assembly (described in detail in Methods and Supplementary Table 3). **c**,**d**, Decomposed kinetics of postmitotic (**c**) and interphase (**d**) assembly from (**b**).

Source data

Extended Data Fig. 6 Kinetic decomposition of the two assembly processes for each nucleoporin.

The fluorescence intensities at non-core (brown) and inner-core (yellow) regions were plotted and fitted with a sequential model of NPC assembly (bold lines) as in Extended Data Fig. 5. Dots represent the average and s.d. of measurements from 15, 20, 13, 14, 22, 22, 19, 24, 14 and 13 cells for Nup107, Seh1, Nup205, Nup93, Nup62, Nup214, Tpr, Nup358, Nup153, and Pom121, respectively. Single confocal slices of cells at 20 min, 40 min, and 60 min after AO are shown. Images were filtered with a median filter (kernel size: 0.25 × 0.25 µm) for presentation purposes. Scale bars, 10 µm.

Source data

Extended Data Fig. 7 Computed parameters for the nucleoporin assembly kinetics and the remarkable difference between postmitotic and interphase assembly.

a, An example of computed quantities (Nup358, interphase assembly). **b**, Illustrations of median assembly time of nucleoporins in postmitotic (left) and interphase (right) assembly superimposed onto a simplified scheme of the NPC. The median assembly time of individual nucleoporins are displayed on a pseudo-colour scale. **c**, Plots of Nup assembly duration along the median assembly time for postmitotic and interphase assembly pathways. The crosses indicate the 95% confidence intervals. The dots are the mean values. Values are listed in Supplementary Table 3. The long straight line shows the result of a linear regression to the mean values. The gray area is the 95% confidence interval of the linear regression. Theoretically, for a sequential assembly mechanism where late steps depend on early steps, the observed ensemble kinetics of a late binding protein must contain the history of all previous events (see Methods for details). Indeed, postmitotic assembly showed a strong positive linear correlation, indicating a sequential assembly mechanism, which for example implies that Nup62 is incorporated into the NPC before Tpr can bind.

Extended Data Fig. 8 The variations of the integrative models for postmitotic assembly pathway.

Detailed views of the best and lower-scoring pathways. The side views of the 3 spokes are shown. The NE surface is indicated by wireframe model (grey). The uncertainty of each Nup localization is indicated by the density of the corresponding color: The Y-complex (green), isolated fraction of Nup155 (orange), Nup93-Nup188-Nup155 and Nup205-Nup93-Nup155 complexes (blue), and Nup62-Nup58-Nup54 complex (purple). The posterior model likelihood for each pathway is indicated (see Methods for details). The variations in the structural models for the intermediates at 5 and 6 min after anaphase onset would be due to their lower protein density (lower signal-to-noise ratio) in the electron tomography images.

Extended Data Fig. 9 Nup155 assembles at a similar timing to Nup62 in postmitotic assembly pathway.

a,**b**, Immunofluorescence with an anti-Nup155 antibody and a GFP-Nanobody. Cells at anaphase and telophase were imaged by confocal microscopy. mEGFP-Nup107 (**a**) and Nup62-mEGFP (**b**) genome-edited

cells were used. The cell nuclei were stained with SiR-DNA. Scale bars, 10 μm.

Extended Data Fig. 10 Postmitotic assembly kinetics of Nup188 is similar to other inner ring components.

a, Digital PCR analysis of Nup188-mEGFP genome-edited cells for assessing the homozygous tagging. Allele copy numbers of GFP integrations and on-target GFP integrations are quantified. The data are from 2 and 3 measurements for wild-type and Nup188-mEGFP genome-edited HeLa cells. Error bars represent the s.d. of the mean. **b**, Immunoblot analysis of Nup188-mEGFP cells using antibodies against Nup188 and γ-Tubulin. Uncropped immunoblot data are shown in Supplementary Fig. 1. **c**, Calculated copy number of Nup188 per nuclear pore as in Fig. 1. The plot is from 20 cells. The median is depicted as a line. **d**, Time-lapse 3D imaging of Nup188-mEGFP genome-edited cell. The cell nuclei were stained with silicon–rhodamine (SiR)-DNA. Single confocal sections of SiR and GFP channels are shown. Images were filtered with a median filter (kernel size: 0.25 × 0.25 μm). Scale bar, 10 μm. Time after AO is indicated. **e**, Postmitotic assembly kinetics of Nup188 was decomposed as in Extended Data Fig. 5. Dots represent the average and s.d. of measurements from 16 cells. The kinetics of Nup107, Nup93 and Nup214 are also shown for comparison.

Source data

Supplementary information

Supplementary Figures

Reporting Summary

Peer Review File

Supplementary Table 1

Concentration of nucleoporins (nM). Concentration of mEGFP-tagged nucleoporins in interphase and metaphase. Data represent mean ± S.D. from the number of cells indicated in parentheses.

Supplementary Table 2

gRNA sequences used for genome editing. The first sequence is the antisense gRNA-binding site, the second one is the sense gRNA-binding site. For Nup188, Alt-R S.p. HiFi Cas9 Nuclease V3 was used that requires only single gRNA.

Supplementary Table 3

Parameter of the mathematical model. The number in brackets represents the 95% confidence interval as estimated using profile likelihood. *The fraction of postmitotic assembly in the non-core and core region represent the average for all Nups. See Methods. The assembly in the core region is delayed by $d_c = 2$ min.

Source data

Source Data Fig. 1

Source Data Fig. 3

Source Data Fig. 4

Source Data Extended Data Fig. 2

Source Data Extended Data Fig. 3

Source Data Extended Data Fig. 4

Source Data Extended Data Fig. 5

[**Source Data Extended Data Fig. 6**]

[**Source Data Extended Data Fig. 10**]

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Otsuka, S., Tempkin, J.O.B., Zhang, W. *et al.* A quantitative map of nuclear pore assembly reveals two distinct mechanisms. *Nature* **613**, 575–581 (2023). https://doi.org/10.1038/s41586-022-05528-w

- Received: 18 May 2021
- Accepted: 04 November 2022
- Published: 04 January 2023
- Issue Date: 19 January 2023

- DOI: https://doi.org/10.1038/s41586-022-05528-w

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

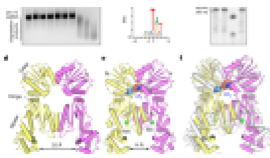
Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

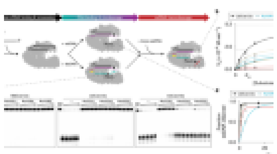
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05528-w



The Card1 nuclease provides defence during type III CRISPR immunity

18 January 2021

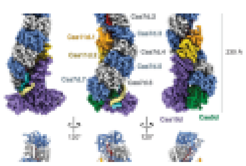
Jakob T. Rostøl, Wei Xie, … Luciano A. Marraffini



Broad-spectrum enzymatic inhibition of CRISPR-Cas12a

01 April 2019

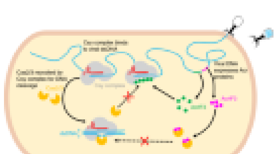
Gavin J. Knott, Brittney W. Thornton, … Jennifer A. Doudna



Structural rearrangements allow nucleic acid discrimination by type I-D Cascade

20 May 2022

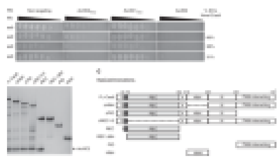
Evan A. Schwartz, Tess M. McBride, … David W. Taylor



[AcrIF5 specifically targets DNA-bound CRISPR-Cas surveillance complex for inhibition]

17 March 2022

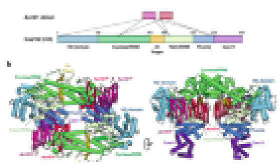
Yongchao Xie, Laixing Zhang, … Yue Feng



[Inhibition of CRISPR-Cas9 ribonucleoprotein complex assembly by anti-CRISPR AcrIIC2]

26 June 2019

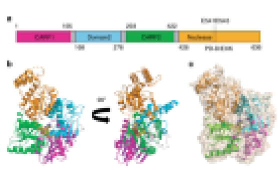
Annoj Thavalingam, Zhi Cheng, … Karen L. Maxwell



[Structural basis for inhibition of an archaeal CRISPR–Cas type I-D large subunit by an anti-CRISPR protein]

25 November 2020

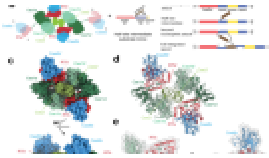
M. Cemre Manav, Lan B. Van, … Ditlev E. Brodersen



[Structure and mechanism of a Type III CRISPR defence DNA nuclease activated by cyclic oligoadenylate]

24 January 2020

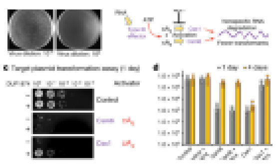
Stephen A. McMahon, Wenlong Zhu, … Tracey M. Gloster



[Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex]

06 May 2021

Joy Y. Wang, Christopher M. Hoel, … Jennifer A. Doudna



[An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity]

15 January 2020

Januka S. Athukoralage, Stephen A. McMahon, … Malcolm F. White

- Article
- Open Access
- Published: 04 January 2023

RNA targeting unleashes indiscriminate nuclease activity of CRISPR–Cas12a2

- Jack P. K. Bravo,
- Thomson Hallmark,
- Bronson Naegle,
- Chase L. Beisel,
- Ryan N. Jackson &
- …
- David W. Taylor

Show authors

Nature **volume 613**, pages 582–587 (2023)

- 14k Accesses
- 1 Citations
- 300 Altmetric
- Metrics details

Abstract

Cas12a2 is a CRISPR-associated nuclease that performs RNA-guided, sequence-nonspecific degradation of single-stranded RNA, single-stranded DNA and double-stranded DNA following recognition of a complementary RNA target, culminating in abortive infection[1]. Here we report structures of Cas12a2 in binary, ternary and quaternary complexes to reveal a complete activation pathway. Our structures reveal that Cas12a2 is autoinhibited until binding a cognate RNA target, which exposes the RuvC active site within a large, positively charged cleft. Double-stranded DNA substrates are captured through duplex distortion and local melting, stabilized by pairs of 'aromatic clamp' residues that are crucial for double-stranded DNA degradation and in vivo immune system function. Our work provides a structural basis for this mechanism of abortive infection to achieve population-level immunity, which can be leveraged to create rational mutants that degrade a spectrum of collateral substrates.

Main

Prokaryotic adaptive immunity typically uses CRISPR–Cas systems to target and degrade foreign genetic elements, including phage and plasmids[2,3]. However, it was recently discovered that Cas12a2 from *Sulfuricurvum* sp. PC08-66 instead relies on abortive infection—that is, dormancy or cell death in response to the presence of an invader—to achieve population-level immunity, preventing the replication and transmission of plasmids[1].

Although Cas12a2 sometimes co-occurs with Cas12a systems in bacteria and can use Cas12a CRISPR RNA (crRNA), Cas12a2 recognizes an RNA target strand with a suitable protospacer-flanking sequence (PFS; for example, 5′-GAAAG-3′) rather than the double-stranded (ds)DNA target of Cas12a[1,4,5]. Furthermore, Cas12a2 is immune to the effects of many anti-CRISPR proteins that target Cas12a, and aside from a conserved RuvC nuclease domain and pre-crRNA processing region, Cas12a and Cas12a2 sequences bear little resemblance to one another (about 10–20% sequence identity). Notably, Cas12a2 lacks a Nuc domain (involved in DNA target strand loading), but instead contains a zinc ribbon (ZR), and it contains a unique insertion domain in place of the Cas12a bridge helix.

Unlike many of the recently characterized abortive infection systems[6,7,8,9,10,11,12,13,14], Cas12a2 does not rely on the production of secondary messengers to achieve antiphage immunity. Instead, Cas12a2 activation induces robust, nonspecific cleavage of single-stranded (ss)RNA, ssDNA and dsDNA[1] in *trans*. This mechanism is unique to Cas12a2, although the molecular basis for collateral nucleic acid degradation is unknown.

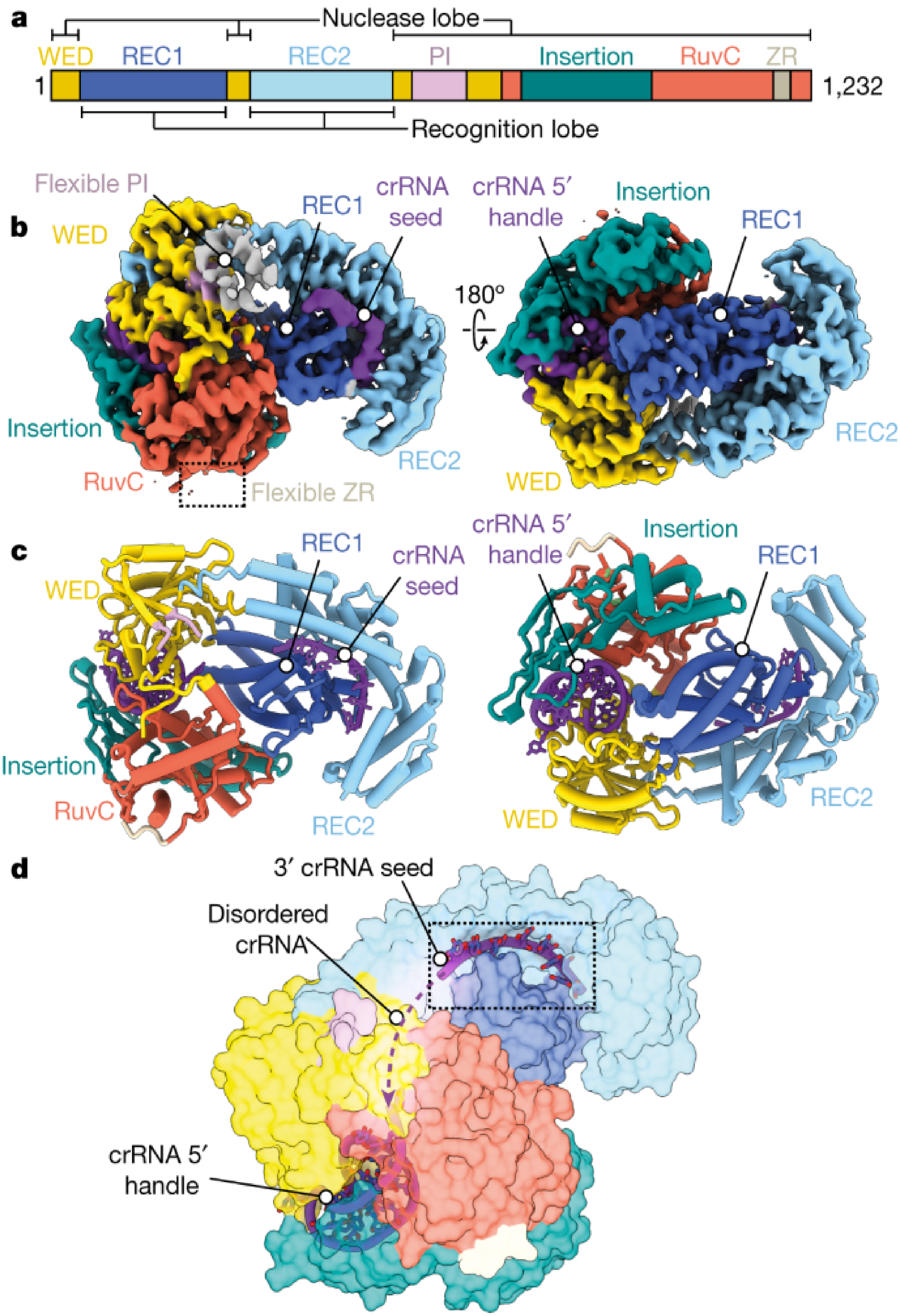
To understand the unique mechanisms of activation, substrate capture and indiscriminate nuclease activity underlying the function of Cas12a2, we carried out biochemical, structural and in vivo analyses, including determining cryo-electron microscopy (cryo-EM) structures of autoinhibited Cas12a2–crRNA (binary complex) associated with an RNA target (ternary complex) and bound to both an RNA target and a dsDNA collateral substrate mimetic (quaternary complex).

Structure of Cas12a2 binary complex

To gain insights into the function of Cas12a2, we first purified a binary complex consisting of catalytically active Cas12a2 and a mature crRNA and determined the structure using cryo-EM to a global resolution of 3.2 Å (Extended Data Table 1). The quality of the map allowed de novo modelling of most of Cas12a2, aside from the flexible PFS-interacting (PI) and ZR domains.

Cas12a2 adopts a bilobed architecture (Fig. 1), with a recognition (REC) lobe comprising the REC1 and REC2 domains and the nuclease (NUC) lobe consisting of the PI, wedge (WED), RuvC nuclease, ZR and Cas12a2-specific insertion domains. The overall architecture of the binary complex resembles an oyster, as opposed to the triangular 'sea conch' shape of Cas12a[15].

Fig. 1: Cas12a2 binary complex resembles an oyster and orders the crRNA.



a
Nuclease lobe
WED
REC1
REC2
PI
Insertion
RuvC
ZR
1
1,232
Recognition lobe
b
Flexible PI
WED
REC1
crRNA
seed
crRNA 5′
handle
Insertion
REC1
180°
Insertion
RuvC
Flexible ZR
REC2
WED
REC2
c
WED
REC1
crRNA
seed
crRNA 5′
handle
Insertion
REC1
Insertion
RuvC
REC2
WED
REC2
d
3′ crRNA seed
Disordered
crRNA
crRNA 5′
handle

a, Domain organization of Cas12a2. **b**, Cryo-EM structure of Cas12a2 binary complex coloured by structural domain as in **a**. **c**, Atomic model of Cas12a2 binary complex. **d**, Putative 3′ seed region of crRNA. Seven bases from the 3′ end are ordered, and bases are solvent-exposed, probably acting as a seed for target RNA binding.

Although Cas12a and Cas12a2 share low (10–20%) sequence similarity, comparison of their WED and RuvC domains shows a high degree of structural similarity (root mean square deviation of 1.073 Å across 120 equivalent residues with FnCas12a; PDB ID 5NG6; Extended Data Fig. 2). Furthermore, the crRNA 5′ stem–loop is in an identical configuration in both complexes, and a loop containing basic residues is similarly positioned to catalyse pre-crRNA maturation. The common 'chassis' formed by these domains provides a structural scaffold that enables the same crRNAs to prime and guide either complex to the same target sequences for their different functions.

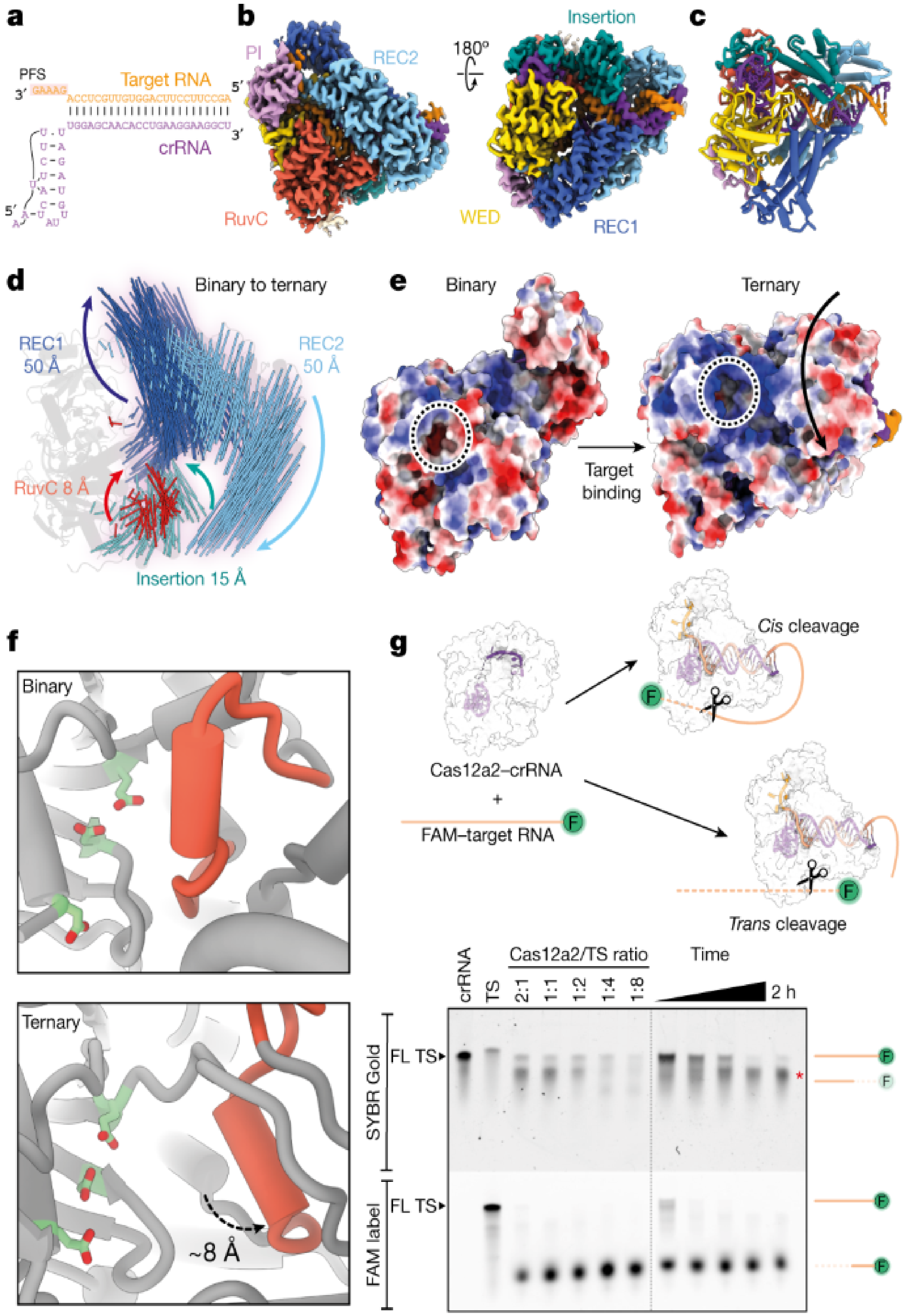
Despite the similar domain organization to Cas12a within the WED and RuvC domains, Cas12a2 has a unique α-helical REC lobe, with no known structural homologues. The differences in the structural organization of the REC lobe probably allow Cas12a2 to escape targeting by many anti-CRISPR proteins that can efficiently shut down Cas12a[1] (Extended Data Fig. 3). Of note, seven nucleotides of pre-ordered crRNA sit at the interface between REC1 and REC2 in a conformation in which bases are solvent-exposed and primed for target recognition (Fig. 1d). This region is towards the 3′ end of the crRNA, and the intervening sequence between the 5′ crRNA stem–loop and this pre-ordered guide is disordered in our structure, probably owing to flexibility. This is in contrast to the case for Cas12a, which has a well-described pre-ordered crRNA immediately flanking the 5′ stem–loop. In Cas12a, this region is highly sensitive to mismatches because it initiates R-loop formation following recognition of the protospacer adjacent motif (PAM) and is considered a seed region[16]. By contrast, Cas12a2 is insensitive to single mismatches within the entirety of the crRNA but has reduced in vivo activity when truncated on the 3′ end[1]. Our structure suggests that Cas12a2–crRNA–target strand (TS) duplex formation may initiate and propagate from the 3′ end of the crRNA, enabling Cas12a2 to

target phage that have escaped surveillance by Cas12a through mutagenesis of the PAM or 5′ seed regions.

RNA target binding activates Cas12a2

To understand how RNA target recognition is distinct from other Cas12 family nucleases, we next investigated how RNA target binding activates Cas12a2. We determined a 2.9-Å-resolution cryo-EM structure of a ternary complex consisting of Cas12a2, crRNA and a target ssRNA containing the non-self PFS 5′-GAAAG-3′ (Fig. 2).

Fig. 2: Target binding leads to large-scale structural arrangements for activation.



a
PFS
Target RNA
3′ GAAAG ACCUCGUUGUGGACUUCCUUCCGA 5′
UGGAGCAACACCUGAAGGAAGGCU 3′
crRNA
5′
b
PI
REC2
180°
Insertion
RuvC
WED
REC1
c
d
Binary to ternary
REC1 50 Å
REC2 50 Å
RuvC 8 Å
Insertion 15 Å
e
Binary
Ternary
Target binding
f
Binary
Ternary
~8 Å
g
Cis cleavage
Cas12a2–crRNA
+
FAM–target RNA
Trans cleavage
crRNA
TS
Cas12a2/TS ratio
2:1
1:1
1:2
1:4
1:8
Time
2 h
SYBR Gold
FL TS
FAM label
FL TS

a, Schematic of crRNA–TS duplex. PFS on target RNA is highlighted. **b**, Cryo-EM structure of Cas12a2 ternary complex. **c**, Atomic model of Cas12a2 ternary complex. **d**, Motion vector map showing conformational changes of Cas12a2 induced following ternary complex formation. Binary complex model shown as grey cartoon. Conformational changes induced by target RNA binding are also shown in Supplementary Video 1. **e**, Surface electrostatic potential of binary and ternary complex, showing how active site (dotted circle) becomes exposed following ternary complex formation. This is accompanied by the formation of a large positively charged groove adjacent to the activate site. **f**, Displacement of RuvC gating helix (red) by about 8 Å following ternary complex formation exposes active site residues (light green). **g**, Target protection revealed by SYBR staining, showing only peripheral cleavage of target RNA when binary complex is in molar excess, and total degradation when target is in excess. The observed protection persisted for 2 h with excess binary complex. Representative of three independent experiments with similar results. For gel source data, see Supplementary Fig. 1. FL, full length.

The 22-base-pair A-form crRNA–target RNA duplex runs through the centre of the complex. At the 5′ end of the crRNA guide, the duplex splits with the crRNA 5′ stem–loop wedged between the RuvC and WED domains and the 3′ PFS end of the target RNA is gripped by the PI domain, which has now become ordered (Fig. 2b). Each of the five nucleotides of the PFS make specific base contacts with residues within the PI domain, including hydrogen bonding and π–π stacking (Extended Data Fig. 4). These contacts stabilize the otherwise flexible PI domain, allowing Cas12a2 to distinguish self (that is, complementary to the crRNA 5′ handle) from non-self target RNA on the basis of the PFS. Removal of the PI domain had no effect on the overall structure of Cas12a2, but prevented activation of nuclease activity (Extended Data Fig. 4). Notably, this is a completely distinct mechanism of self-versus-non-self discrimination from that of Cas13 and several type III effector complexes, for which nuclease activity is inhibited by additional complementarity with the 5′ crRNA tag region of the crRNA[17,18,19]. By contrast, Cas12a2 is exclusively activated following recognition of an appropriate PFS sequence.

Cas12a2 is unable to degrade nucleic acids in the absence of a suitable target RNA[1] (Extended Data Fig. 4). Superposition of the binary and ternary

complexes reveals substantial conformational changes localized to the REC1 and REC2 domains, whereas the NUC lobe remains predominantly static (Fig. 2d). REC1 and REC2 are both displaced by up to about 50 Å and move in different directions, creating a central channel, which accommodates the crRNA–TS duplex. The insertion domain moves by up to about 15 Å, but these changes are exclusively localized to the carboxy-terminal half of the domain (residues 938–1,030). The amino-terminal half (residues 870–937), which makes numerous contacts with the crRNA 5′ stem–loop, remains static. On the basis of this observation, we propose that the insertion domain acts as a transducer, allowing allosteric communication between the REC and NUC lobes in response to target RNA binding (Extended Data Fig. 5).

Inspection of the RuvC active site in the autoinhibited binary complex reveals that the catalytic triad (D848, E1063 and D1213) is buried within a solvent-excluded pocket (Fig. 2e). Strikingly, the conformational changes that accompany target RNA binding create a 25-Å-wide positively charged groove that exposes the active site (Fig. 2e). This groove is of sufficient size to accommodate both single- and double-stranded nucleic acids. Although other Cas12 proteins undergo conformational changes following crRNA hybridization (up to about 25 Å for Cas12a[20] and Cas12j[21], but more typically up to about 10 Å (ref. [22])), the approximately 50-Å conformational rearrangements we observe for Cas12a2 are considerably larger, highlighting the distinct activation mechanism of Cas12a2 (Supplementary Video 1).

Access to the RuvC catalytic triad is also mediated by the approximately 8-Å shift of a lid helix, which contributes to the change in active site solvent exposure (Fig. 2f). This is akin to the lid loop or helix that gates active site exposure reported for other Cas12 endonucleases[23,24].

Once activated by target binding, the lid of these endonucleases remains ‘open’, enabling ssDNA cleavage in *trans*. However, in previously reported Cas12a structures (including FnCas12a2, AsCas12a and LbCas12a), the RuvC active site is buried owing to the presence of the Nuc domain[25,26,27,28]. In structures of catalytically dead Cas12b and Cas12i, bystander ssDNA bound in *trans* is tightly interwoven to sit within the active site[22,23] (Extended Data Fig. 6). This is in contrast to the highly accessible Cas12a2 RuvC active site in the ternary complex, providing a structural

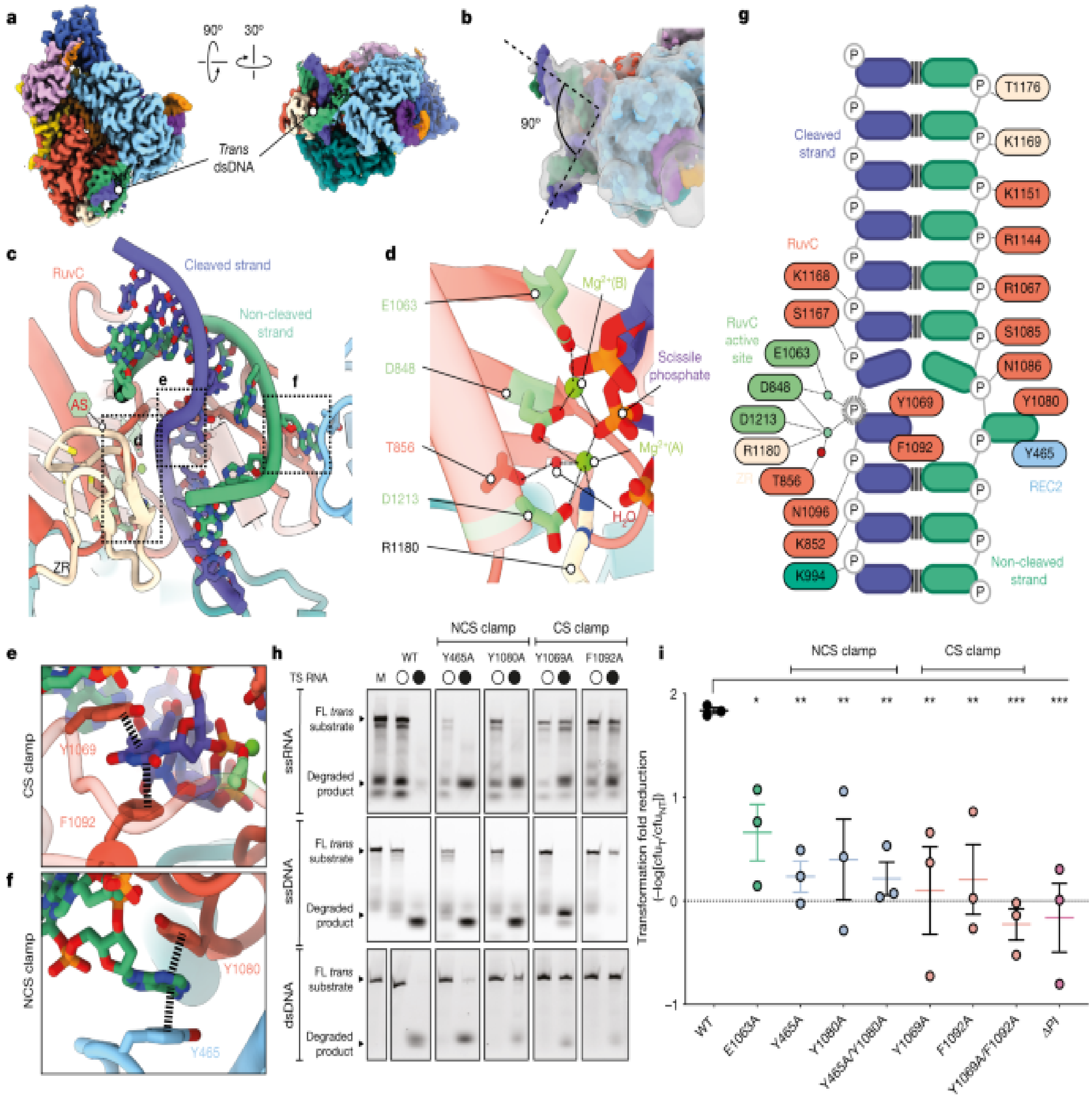
basis for efficient cleavage of a wide range of substrates in *trans*. The lack of a Nuc domain and the presence of a highly exposed RuvC active site in the ternary structure thus explain why Cas12a2 collateral nuclease activation results in an abortive infection phenotype[1] whereas Cas12a collateral ssDNase activity does not play a role in bacterial immunity[29]. The Nuc domain may act as a physical barrier to the RuvC active site in other Cas12 enzymes (including Cas12a), limiting cleavage in *trans*.

Unlike in Cas12a, the path followed by the target RNA–crRNA duplex completely circumvents the RuvC active site, suggesting a mechanism for RNA degradation in *trans*. To test this, we incubated the Cas12a2 binary complex with fluorescently labelled target RNA at a range of molar ratios and analysed RNA cleavage. The 5′ fluorescein label was consistently trimmed owing to the approximately 20-nucleotide flexible RNA sequence extending from the spacer (Fig. 2g). SYBR Gold staining revealed that even though the extended single-stranded 5′ end of the target RNA was trimmed, the target RNA otherwise remained intact and was protected from degradation with a molar equivalence or excess of Cas12a2. This is distinct from the cleavage mechanisms of other Cas12 nucleases that achieve antiphage immunity through cleavage in *cis*[16,29], and is reminiscent of Cas13 RNase activity in *cis* and *trans*, for which the hybridized region of the target RNA remains intact[18].

Collateral dsDNA binding through duplex contortion

We next sought to visualize how Cas12a2 can accommodate and degrade nucleic acid duplexes. To this end, we determined a 2.7-Å-resolution structure of crRNA-guided Cas12a2 bound both to an activating target RNA and to collateral phosphorothioate dsDNA substrate (Fig. 3). The RuvC active site and 11 of 20 base pairs of the DNA duplex were well resolved, whereas the flexible DNA ends are visible only at lower density thresholds.

Fig. 3: Cas12a2 binds and clamps duplex DNA.



a, Cryo-EM structure of Cas12a2 quaternary complex. *Trans*-dsDNA is shown as slate blue and sea green. **b**, Atomic model of Cas12a2 quaternary complex, with 5 Å low-pass filtered map shown in transparent grey. To highlight the 90º kink in the collateral duplex, two linear ideal B-form dsDNA models have been rigid-body fitted into the map. **c**, dsDNA situated within active site. **d**, Close-up view of Cas12a2 active site. **e**, CS held in place through aromatic clamp. **f**, NCS held in place through aromatic clamp. **g**, Schematic of interactions between Cas12a2 and the collateral dsDNA substrate. CS scissile phosphate is denoted by dashed outline. **h**, Cleavage of ssRNA (top), ssDNA (middle) and dsDNA (bottom) by Cas12a2, and aromatic clamp mutants. M, size marker (intact substrate). Representative of

three independent experiments with similar results. **i**, Alterations to essential residues resulted in loss of the ability of Cas12a2 to clear plasmid (that is, lower transformation fold reduction, calculated as $-\log_{10}[\mathrm{cfu_T}/\mathrm{cfu_{NT}}]$, in which $\mathrm{cfu_T}$ and $\mathrm{cfu_{NT}}$ represent the number of colony-forming units for target and non-target plasmids, respectively). Significance between WT and mutant SuCas12a2 was determined by two-sided Student *t*-test. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Experiments were carried out in triplicate, and error bars correspond to the mean and standard error. For gel source data, see Supplementary Fig. 1.

Source Data

In our structure, the dsDNA duplex is sharply bent by about 90º (Fig. 3b), resulting in duplex distortion and local melting of two base pairs in the immediate vicinity of the RuvC active site (Fig. 3c), enabling the positioning of the scissile phosphate adjacent to the RuvC catalytic triad (Fig. 3d). We designate the DNA strand within the RuvC active site as the cleaved strand (CS) and its complement DNA strand as the non-cleaved strand (NCS) to differentiate from target and non-target strand nomenclature used to describe the strands of dsDNA that are specifically targeted with a crRNA guide (for example, Cas12a or Cas9). As both DNA strands contained non-hydrolysable phosphorothioate modifications, we could visualize the pre-hydrolysis RuvC active site state, including two Mg^{2+} ions and a putative activating water adjacent to one of the ions (designated Mg^{2+}(A); Fig. 3d and Extended Data Fig. 7).

Adjacent to the RuvC active site, both the CS and NCS are stabilized by a large network of nonspecific interactions with Cas12a2 (Fig. 3c,g). This hub of contacts with both duplex ends induces duplex bending and local melting. The melted bases are subsequently captured by two pairs of 'aromatic clamps' (Y465 and Y1080, Y1069 and F1092, respectively) that each hold a single DNA base through π–π stacking, preventing rehybridization (Fig. 3e,f). We confirmed this result through in vitro collateral nuclease assays (Fig. 3h): whereas wild-type (WT) Cas12a2 degraded ssRNA, ssDNA and dsDNA in *trans* when activated with complementary target RNA, the NCS clamp alterations Y465A and Y1080A seemed to reduce duplex cleavage, while having no effect on ssRNase or ssDNase nuclease activity. This

indicates that unwinding by the NCS clamp is critical for nuclease activity of a DNA duplex. CS clamp alterations Y1069A and F1092A abrogate dsDNase activity as expected. Notably, whereas both CS clamp alterations prevent ssRNase activity, only F1092A blocks ssDNase activity (Fig. 3h). To better understand how these mutants would affect in vivo activity, we further tested these point mutations on Cas12a2 supercoiled plasmid DNA cleavage activity, and found that all had severely reduced activity (Extended Data Fig. 8). Of note, the NCS(Y465A) mutant that can cleave a linear FAM-labelled substrate (Fig. 3h) is also able to nick and/or linearize supercoiled plasmid, but is unable to degrade DNA in a timescale similar to that of the WT, providing a rationale for the in vivo effects of this mutant.

As Y1069A preserves ssDNase activity but prevents ssRNase and dsDNase activity, this mutant may enable development of a molecular biosensor that degrades a fluorescence reporter ssDNA following recognition of a complementary ssRNA, enabling sensitive detection of RNA without target depletion as is the case in Cas13 reporter systems.

As NCS clamp alterations ablate dsDNase activity but have no effect on single-stranded collateral nuclease activity, we tested these separation-of-function mutants in vivo. As expected, CS alterations and PI domain truncation blocked Cas12a2 activity (Fig. 3i and Extended Data Fig. 9). Alteration of NCS clamp residues individually, and simultaneously, also abrogated the ability of Cas12a2 to clear target plasmid in vivo, providing direct evidence that aromatic clamp-mediated dsDNA duplex melting is essential for Cas12a2 activity. These data also suggest that dsDNA duplex degradation is the driving force behind Cas12a2-mediated immunity, as mutants that retained ssRNase and ssDNase activities were not sufficient to provide immunity.

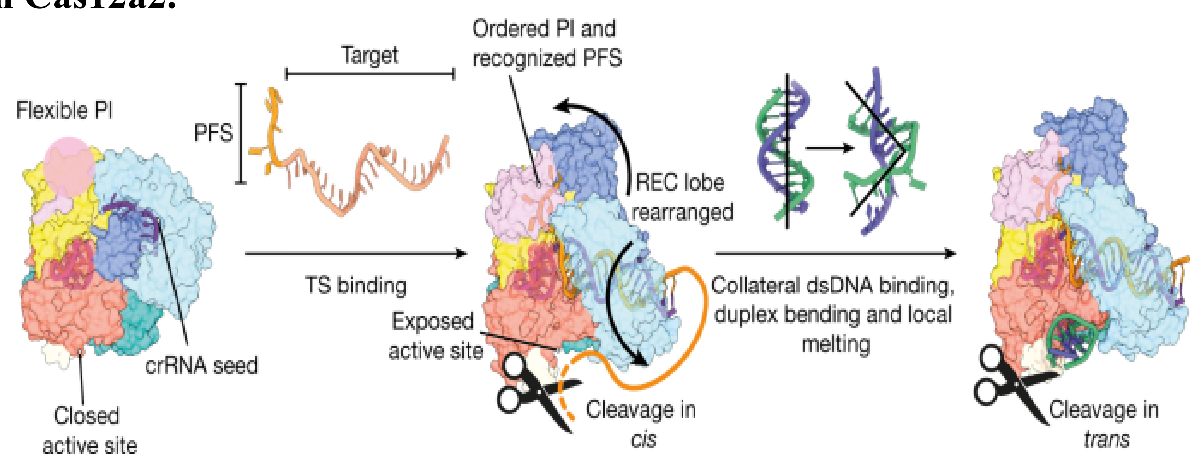
Collectively these data show that Cas12a2 mediates abortive infection through a unique mechanism of dsDNA cleavage, distinct from indiscriminate RNA cleavage-induced abortive infection in other RNA-targeting CRISPR–Cas systems[13,30]. The catalytic mechanism of Cas12a2 is consistent with that of other RuvC endonucleases[21,31,32,33], for which protein-induced structural tension of the DNA facilitates proper scissile phosphate coordination. However, compared to the available RuvC-

containing Cas9 and Cas12 structures, the NCS aromatic clamps provide a unique strategy to cleave duplexed nucleic acids.

Discussion

Our results support a detailed mechanism of Cas12a2 in antiphage defence. Hybridization of the crRNA to PFS-containing RNA targets drives major conformational changes in Cas12a2, exposing the RuvC active site and alleviating autoinhibition (Fig. 4). In the active complex, the RuvC domain is located within an approximately 30-Å-wide positively charged groove that is sufficiently large to accommodate duplexed nucleic acids. Nonspecific electrostatic interactions facilitate collateral substrate capture, accompanied by duplex distortion and local base-pair melting. The melted bases are stabilized by two pairs of aromatic clamps that enable appropriate positioning of single nucleic acid strands within the RuvC active site. This multiple-turnover DNA nicking culminates in dsDNA degradation, distinct from the single-turnover dsDNA cleavage of Cas12a or Cas9 (ref. [16,34]) and enables robust and widespread DNA destruction in vivo.

Fig. 4: Mechanism of RNA target-activated collateral nuclease activity in Cas12a2.



In the binary complex (Cas12a2–crRNA), the active site is occluded. The 3′ end of the crRNA is a pre-ordered seed, and the PI domain is flexible. Binding of a complementary RNA target strand (TS) containing a PFS triggers significant conformational changes within the REC lobe, exposing the RuvC active site to facilitate TS RNA trimming in *cis* and nuclease activity in *trans*. This enables binding of dsDNA in *trans*, resulting in

duplex kinking and local melting. The duplex is held open through the action of two pairs of aromatic clamp residues (Fig. 3), enabling nicking by RuvC.

Although Cas12a and Cas12a2 can use the same crRNA owing to their similar WED and RuvC domains, multiple structural divergences confer disparate target preferences and biochemical activities, and enable Cas12a2 to evade inhibition by multiple Cas12a-targeting anti-CRISPR proteins. Notably, the lack of a Nuc domain (also referred to as target loading domain, or target nucleic acid binding domain[35,36]) may contribute to the inability of Cas12a2 to directly bind dsDNA targets through the formation of an R-loop, conferring specificity for ssRNA target binding. Additionally, the lack of a Cas12a2 Nuc domain increases the accessibility to the RuvC nuclease active site, enabling rapid duplex capture and cleavage in *trans*. Previous structures of Cas12b and Cas12i bound to bystander ssDNA have shown that *trans* ssDNA substrates[22,23] must follow a tortuous, narrow path to reach the RuvC active site, which is incompatible with rigid dsDNA substrates (Extended Data Fig. 6). This is in contrast to the highly exposed Cas12a2 RuvC active site, providing a structural mechanism for dsDNA cleavage in *trans*.

While the mechanism of Cas12a2 is reminiscent of RNA targeting by Cas13, there are several distinct differences. Target binding by Cas13 activates nonspecific RNA degradation both in *cis* and in *trans*, resulting in persistent nuclease activation as the phage genome continues to produce target transcripts[30,37,38]. RNA cleavage can occur in *cis* if the target is of suitable length to fully hybridize with the crRNA and reach the distal active site opposite the seed region[39]. Similarly, Cas12a2 can trim target RNA in *cis*, but the hybridized target strand remains intact, enabling uninterrupted Cas12a2 activation (Extended Data Fig. 10). Cas13 systems therefore act as sentinels for viral RNA, whereas Cas12a2 represents the self-destruct button, conferring population-level antiphage defence through abortive infection. The consequences of Cas12a2 activation may be mitigated through protein degradation and turnover coupled with the removal of phage transcript-encoding DNA by other defence systems (for example, restriction–modification systems and DNA-targeting CRISPR–Cas systems)[40].

Our discovery of the Cas12a2(Y1069A) point mutant that can cleave ssDNA but not dsDNA or ssRNA provides a blueprint for an RNA sensor with collateral activity that does not destroy the RNA target it has been programmed to detect or get distracted by the excess of other RNAs in the sample, which might enable the development of a highly sensitive RNA diagnostic.

Methods

Mutagenesis

Strains, plasmids, primers and nucleic acid substrates can be found in Supplementary Table 1. The mutant constructs used in this paper were prepared by point-directed mutagenesis using the Q5 Site-Directed Mutagenesis Kit (NEB). Primers were designed and annealing temperatures were determined using the NEBase Changer tool. All primers were ordered through IDT. Plasmid ligation was achieved with KLD Enzyme Mix (NEB) and completed plasmids were sequenced to verify correct alterations (Plasmidsaurus).

Expression and purification

N-terminal hexa-his-tagged Cas12a2 and various mutant plasmids were transformed into chemically competent *Escherichia coli* NiCo21 cells (NEB). A single colony from transformation was selected for starter culture in 20 ml LB medium grown at 37 °C overnight (16–18 h). Each starter culture was used to inoculate 1.0 l of TB medium, and then cultures were grown up to an optical density at 600 nm of 0.6 at 37 °C. Cultures were cooled in ice for 15 min before being grown at 18 °C for a further 16–18 h followed by collection by centrifugation. Cell pellets were stored at −80 °C or used immediately for protein purification.

Cell pellets were resuspended in lysis buffer (25 mM Tris pH 7.2, 500 mM NaCl, 2 mM $MgCl_2$, 10 mM imidazole, 10% glycerol) treated with protease inhibitors (2 μg ml^{-1} aprotinin, 10 μM leupeptin, 0.2 mM AEBSF, 1.0 μg ml^{-1} pepstatin) and 1 mg ml^{-1} lysozyme and incubated for 30 min on ice

with shaking. Cells were lysed by sonication and clarified by centrifugation. Clarified lysate was batch-bound to nickel resin for 30 min at 4 °C and then allowed to flow through. Lysate was then passed again over nickel resin twice. Nickel resin was washed with nickel wash buffer (25 mM Tris pH 7.2, 2 M NaCl, 2 mM $MgCl_2$, 10 mM imidazole, 10% glycerol) and eluted with nickel elution buffer (25 mM Tris pH 7.2, 500 mM NaCl, 2 mM $MgCl_2$, 250 mM imidazole, 10% glycerol). Nickel elutions were desalted into low-salt buffer (25 mM Tris pH 7.2, 50 mM NaCl, 2 mM $MgCl_2$, 10% glycerol) using either HiPrep 26/10 desalting column (Cytiva) or Pd10 Sephadex G25M Columns (Cytiva) depending on purification size. Protein samples were then loaded over an ion-exchange column (HiTrap SP HP or HiTrap Q HP column (Cytiva) depending on the isoelectric point of the construct) using low-salt buffer and eluted with a gradient of high-salt buffer (25 mM Tris pH 7.2, 1.0 M NaCl, 2 mM $MgCl_2$, 10% glycerol).

Peak elutions were then pooled and concentrated to about 1.0 ml. During this process, protein to be used for biochemistry was desalted by refilling the concentrator with low-salt buffer twice to exchange out the high-salt buffer. Concentrated protein for cryo-EM was loaded over a HiLoad 26/600 Superdex 200 pg column (GE Healthcare) using SEC buffer (25 mM HEPES pH 7.2, 150 mM NaCl, 2 mM $MgCl_2$, 5% glycerol). Peak fractions were then pooled and concentrated again. Concentrated protein was either flash frozen in liquid nitrogen or used in complex formation for cryo-EM.

Complex formation for cryo-EM

Before crRNA was added to Cas12a2, RNA was incubated at 65 °C for 3 min followed by cooling 1 °C min^{-1} to room temperature. Binary complex was formed for cryo-EM by combining protein and synthetic crRNA in a 1:1.2 molar ratio in SEC buffer (25 mM HEPES pH 7.2, 150 mM NaCl, 2 mM $MgCl_2$, 5% glycerol) and incubating at 24 °C for 10 min. Unbound crRNA was then separated from the binary complex over a Superdex 200 10/300 increase GL sizing column (Cytiva) into Cryo-EM Buffer (12.5 mM HEPES pH 7.2, 150 mM NaCl, 2 mM $MgCl_2$). Eluted protein was concentrated to 30 μM in a 100-kDa MWCO spin concentrator (Corning) and flash frozen in liquid nitrogen.

Far-UV circular dichroism spectroscopy

Protein samples of Cas12a2 mutants were prepared at a concentration in the range 0.3–0.5 mg ml^{-1} determined by nanodrop in circular dichroism (CD) buffer (20 mM Tris pH 7.2, 100 mM NaCl). Far-UV CD readings used a Jasco-J1500 spectropolarimeter. The CD spectra were obtained from 260–190 nm using a scanning speed of 50 nm min^{-1} (with a 2 s response time and accumulation of three scans). Melting curves of Cas12a2(ΔPI) samples (in sealed quartz cuvettes with 0.1 cm path length) were obtained by monitoring the CD signal at 222 nm every 1 °C over a 10–90 °C temperature range, using a temperature ramp of 15 °C h^{-1}. The CD signal was converted to molar ellipticity by Jasco Spectra Manager software.

Plasmid curing assay

Plasmid curing assays were conducted as described previously[1]. In short, immune system plasmids were prepared with Cas12a2 and a 3× CRISPR repeat and transformed into BL21 AI cells by heat-shock transformation. Cells expressing the immune system were then made electrocompetent[41] and immediately transformed by electroporation with 50 ng of either target or non-target plasmid. Transformations were recovered for 18 h in 450 µl LB medium containing 1 mM IPTG, 0.2% l-arabinose and antibiotics for the immune system plasmid. Recovered transformations were then serially diluted in LB medium between 10^1 and 10^6 and spotted in 10-µl drops on LB agar plates containing 1 mM IPTG, 0.2% l-arabinose and antibiotics for both immune system and target or non-target plasmids. Colonies were counted in the highest countable spot in the dilution series and the relative transformation efficiency was calculated between the target and non-target plasmid.

Activation assay

Binary complex of either WT Cas12a2 or Cas12a2(ΔPI) with crRNA was combined with various targets (FL, ΔPFS, Δ5 and Δ10) to final reaction conditions of 600 nM Cas12a2, 720 nM crRNA and 300 nM FAM-labelled target in 1× NEB 3.1 Buffer (50 mM Tris pH 7.9, 100 mM NaCl, 10 mM

$MgCl_2$, 100 µg ml^{-1} BSA). Binary complex was first formed by combining WT Cas12a2 or Cas12a2(ΔPI) with crRNA in a 1:1.2 molar ratio and incubating for 30 min at room temperature with NEB 3.1 buffer as a 2× master mix. Binary complex and target RNAs were then combined to their final reaction concentration and incubated at room temperature for 1.0 h before quenching with 1:1 (v/v) phenol–chloroform pH 4.5 mixed by flicking.

Reactions were quenched with phenol–chloroform pH 4.5 and mixed by flicking followed by spinning down for 30 s. Reaction products were run on a 12% fully denaturing formaldehyde (FDF) polyacrylamide gel electrophoresis (PAGE) gel as described by previously[42] with minor modifications. Loading dye was replaced with 30% glycerol, and gels were run at 50 V for 15 min before increasing the voltage to 150 V.

Target protection assay

Binary complex of Cas12a2 and crRNA was combined with FAM-labelled FL target RNA to a final reaction condition of 600 nM Cas12a2, 720 nM crRNA, 300 nM target RNA in 1× NEB 3.1 buffer (50 mM Tris pH 7.9, 100 mM NaCl, 10 mM $MgCl_2$, 100 µg ml^{-1} BSA). Binary complex was first formed as described in the activation assay. The 2× master mix was then combined with the target RNA and incubated at 37 °C. Samples were taken at time points 5, 15, 30, 60 and 120 min, and quenched in pH 4.5 phenol–chloroform. Samples were then run on a 12% FDF–PAGE gel as described in the activation assay. Completed gels were imaged for FAM fluorescence and then stained with SYBR Gold and imaged again to show unlabelled RNA species.

Effect of ratios on cleavage

FAM-labelled FL target RNA (300 nM) was combined with a range of binary complex concentrations (600, 300, 150, 75 and 37.5 nM) to achieve complex/target ratios of 2:1, 1:1, 1:2, 1:4 and 1:8. Binary complex was first formed by combining Cas12a2 (1,200 nM) with crRNA in a 1:1.2 molar ratio and incubating for 30 min at 37 °C with NEB 3.1 buffer as a 2× master mix. Binary complex was then serially diluted in 2× NEB 3.1 buffer (100

mM Tris pH 7.9, 200 mM NaCl, 20 mM $MgCl_2$, 200 μg ml^{-1} BSA) to form a 2× master mix for each complex/target ratio. Each 2× master mix was combined with target RNA and incubated at 37 °C for 1.0 h followed by quenching with phenol–chloroform pH 4.5. Quenched samples were visualized as described in the activation assay.

***Trans*-cleavage assay**

Reactions containing 600 nM Cas12a2 (WT or mutants) and 720 nM crRNA were combined with 300 nM FL target RNA and 300 nM FAM-labelled non-target RNA, ssDNA or dsDNA. Binary complex was first formed as described previously as a 4× master mix. The master mix was combined with target and non-target substrates and incubated at 37 °C for 1.0 h and quenched with phenol–chloroform pH 4.5. Samples were then separated on a 12% 7 M urea PAGE gel and imaged for FAM fluorescence.

To test the effect of pre-incubation of target with Cas12a2, 100 nM Cas12a2–crRNA complex was incubated for 2 h at room temperature with 200 nM target RNA in NEB 3.1 buffer (50 mM Tris-HCl pH 7.9, 100 mM NaCl, 10 mM $MgCl_2$, 100 μg ml^{-1} BSA). After incubation, Cas12a2/crRNA/target mixture was combined with 1 μM RNAse Alert or DNAse Alert (IDT), and the reactions were allowed to proceed for 60 min at room temperature. These reactions were compared to the same reaction condition with target RNA added simultaneously (no incubation) with RNAse Alert or DNAse Alert. Fluorescent signal was tracked with a Synergy H4 plate reader (BioTek) and data were plotted in GraphPad Prism.

Plasmid cleavage assay

Plasmid cleavage reactions were prepared by combining 14 nM Cas12a2 (or mutants) with 14 nM crRNA and 25 nM target RNA in NEB 3.1 buffer (50 mM Tris-HCl pH 7.9, 100 mM NaCl, 10 mM $MgCl_2$, 100 μg ml^{-1} BSA). Protein was preheated at 37 °C for 15 min before the addition of 7 nM supercoiled pUC19 plasmid. Samples were taken at time points of 1, 2, 5, 10, 20, 30, 45 and 60 min and quenched in pH 8.0 phenol–chloroform. Quenched reactions were mixed by flicking followed by centrifugation.

Samples were loaded on 1% agarose gels and visualized with ethidium bromide.

Cryo-EM sample preparation and data acquisition and processing

Flash-frozen Cas12a2 binary complex was rapidly thawed. A 4 µl volume of the binary complex was applied to C-flat holey carbon grids (2/2, 400 mesh) that had been plasma-cleaned for 30 s in a Solarus 950 plasma cleaner (Gatan) with a 4:1 ratio of O_2/H_2. Grids were blotted with Vitrobot Mark IV (Thermo Fisher) for 2 s, blot force 4 at 4 °C and 100% humidity, and plunge-frozen in liquid ethane. Data were collected on an FEI Glacios cryo-TEM equipped with a Falcon 4 detector. Data were collected in SerialEM v3.8, with a pixel size of 0.94 Å, a defocus range of −1.5 to −2.5 µm and a total exposure time of 15 s resulting in a total accumulated dose of 40 electrons $Å^{-2}$ that was split into 60 electron event representation fractions. Motion correction, contrast transfer function (CTF) estimation and particle picking was carried out on-the-fly using cryoSPARC Live v4.0.0-privatebeta.2 (ref. [43]). A total of 1,577 videos were collected, of which 1,159 were accepted on the basis of meeting the criteria of a CTF fit of 5 Å or better. All subsequent data processing was carried out in cryoSPARC v3.2 (ref. [44]).

From 987,122 particles picked, 214,647 were selected from a single round of two-dimensional (2D) classification. These particles were subjected to ab initio reconstruction (three classes) followed by heterogeneous refinement resulting in a final subset of 97,470 particles that yielded a 3.46-Å-resolution structure from non-uniform refinement. Re-extraction of this subset of particles in a 320-pixel box size, splitting of particles into 4 exposure groups and carrying out per-group CTF refinement and per-particle defocus optimization as implemented in non-uniform refinement[45] resulted in a 3.2-Å-resolution reconstruction that was used for modelling.

For the ternary complex, a rapidly thawed Cas12a2 binary complex fraction was supplemented with a fourfold excess of heat-annealed (90 °C for 5 min, and rapidly cooled to 4 °C) PFS-containing RNA target and incubated at room temperature (about 25 °C) for 30 min before vitrification, which was carried out in an identical manner to that for the binary complex as described

above. Data were collected using an FEI Titan Krios cryo-electron microscope equipped with a K3 Summit direct electron detector (Gatan). Images were recorded with SerialEM[46] with a pixel size of 0.81 Å. A total accumulated dose of 70 electrons $Å^{-2}$ during a 6-s exposure was fractionated into 80 frames. A total of 6,940 micrographs were collected, of which 6,614 with CTF fits of 5 Å or better were retained. On-the-fly processing was carried out as described above.

A total of 3,515,037 particles were picked, of which 2,212,319 were selected after 2D classification. Multiple rounds of ab initio reconstruction and heterogeneous refinement resulted in a subset of 192,639 particles that were reconstructed to 2.92-Å-resolution using non-uniform refinement as described above. This map was then used for modelling.

For the quaternary complex, ternary complex was prepared as described above, and incubated with heat-annealed phosphothioate dsDNA duplex for 30 min at room temperature. A 2.5 µl volume of complex was applied to C-flat grids (1.2/1.3, 300 mesh) and blotted for 6 s, blot force 0 at 4 °C and 100% humidity before vitrification. Data were collected on an FEI Glacios cryo-TEM equipped with a Falcon 4 detector, as described for the binary complex. A total of 1,755 videos were collected, of which 1,539 had CTF fits of 5 Å or better and were retained for subsequent processing. On-the-fly motion correction, CTF estimation and particle picking were carried out as described above.

A total of 1,692,368 particles were picked, of which 425,770 were retained after a single round of 2D classification. A single round of ab initio reconstruction (3 classes) followed by heterogeneous refinement yielded a subset of 260,958 particles that were reconstructed to 2.97 Å resolution using non-uniform refinement. Additional rounds of ab initio reconstruction and heterogeneous refinement were used to further classify particles, resulting in a final subset of 104,857 particles. Extraction of said particles with a 384-pixel box, splitting particles into 9 exposure groups, and reconstruction using non-uniform refinement with per-group CTF refinement and per-particle defocus optimization resulted in a 2.74-Å-resolution reconstruction that was then used for modelling.

Model building and figure preparation

A Cas12a binary complex (Protein Data Bank (PDB) 5NG6)[20] was rigid body fitted into the Cas12a2 binary complex map. Although most of the model did not correspond to the map, the RuvC and WED domains generally were consistent. However, pairwise blast of the two proteins revealed multiple gaps or inserts for the relative insertions. However, a single 20-residue hairpin of the Cas12a RuvC domain fitted the Cas12a2 map well and had no gaps or insertions in the pairwise blast. This fragment was isolated and rigid body fitted into the Cas12a2 map, and the sequence was mutated to the corresponding region of Cas12a2 using Coot v1.0 (ref. [47]). This was then used as a fiducial to build the rest of the complex de novo using in Coot. Attempts at using AlphaFold2 (AF2)[48] to generate fragments to fit in the map were unsuccessful as adjacent residues within the WED and RuvC domains were separated by protein sequence and the REC1 and REC2 domain boundaries were not obvious from the sequence alone. However, AF2 was used to validate modelling of small structural domains after-the-fact, for which smaller, compact regions of the model were folded using AF2, and fitted into the map, indicating correct modelling. This was particularly useful when the de novo model contained gaps due to local flexibility.

The 5′ crRNA handle was built using the Cas12a 5′ crRNA as a template (PDB 5NG6). In the binary complex, the seven-nucleotide 3′ seed region was modelled de novo as polyU as it was not possible to unambiguously determine nucleotide identity.

Once fully modelled, Isolde v1.4 (ref. [49]) was used to improve the fit of the model to the map, and real-space refinement as implemented within Phenix v1.19 (ref. [50]) was carried out to optimize model geometry.

For the Cas12a2 ternary complex, the RuvC, WED, and part of the insert domains were in the same conformation as in the binary complex. These were rigid body fitted into the ternary complex map. The REC1, REC2 and the C-terminal half of the insertion domain were separately fitted as rigid bodies into the ternary complex map. The PI domain structure was predicted using AF2, and then manually connected to the rest of the model. The

crRNA–target RNA duplex was modelled as ideal A-form RNA within Coot, and manually connected to the 5′ crRNA handle. The target RNA 3′ PFS was modelled de novo. Coot was used to fit in gaps within the model, and Isolde was then used to improve the quality of model before real-space refinement as described above.

For the quaternary complex, the ternary complex structure was rigid body fitted into the map, and then flexibly fitted using Isolde. The ZR domain structure was predicted using AF2, and manually connected to the rest of the model. The dsDNA duplex was modelled de novo, with one strand modelled as polyT and the other as polyA as it was not possible to unambiguously determine nucleotide identity. Mg^{2+} and Zn^{2+} ions and an activating H_2O were modelled manually using the sharpened map. Isolde and real-space refinement were carried out as described above.

All structural figures and videos were generated using ChimeraX v1.0 (refs. [51,52]), apart from the modevectors, which were generated in PyMol v2.5.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The atomic models of Cas12a2 binary, tertiary and quaternary complexes have been deposited into the PDB with accession codes 8D49, 8D48 and 8D4A, and the corresponding maps have been deposited into the Electron Microscopy Data Bank with codes EMD-29178, EMD-27180 and EMD-27179, respectively. Source data are provided with this paper.

References

1. Dmytrenko, O. et al. Cas12a2 elicits abortive infection via RNA-triggered destruction of double-stranded DNA. *Nature* https://doi.org/10.1038/s41586-022-05559-3 (2022).

2. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).

3. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

4. Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).

5. Yan, W. X. et al. Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88–91 (2019).

6. Tal, N. et al. Cyclic CMP and cyclic UMP mediate bacterial immunity against phages. *Cell* **184**, 5728–5739 (2021).

7. Lowey, B. et al. CBASS immunity uses CARF-related effectors to sense 3′–5′- and 2′–5′-linked cyclic oligonucleotide signals and protect bacteria from phage infection. *Cell* **182**, 38–49 (2020).

8. Duncan-Lowey, B. & Kranzusch, P. J. CBASS phage defense and evolution of antiviral nucleotide signaling. *Curr. Opin. Immunol.* **74**, 156–163 (2022).

9. Bernheim, A. et al. Prokaryotic viperins produce diverse antiviral molecules. *Nature* **589**, 120–124 (2021).

10. Ofir, G. et al. Antiviral activity of bacterial TIR domains via immune signalling molecules. *Nature* **600**, 116–120 (2021).

11. Rostøl, J. T. et al. The Card1 nuclease provides defence during type III CRISPR immunity. *Nature* **590**, 624–629 (2021).

12. Niewoehner, O. et al. Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543–548 (2017).

13. Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III

CRISPR-Cas systems. *Science* **357**, 605–609 (2017).

14. Koopal, B. et al. Short prokaryotic Argonaute systems trigger cell death upon detection of invading DNA. *Cell* **185**, 1471–1486 (2022).

15. Stella, S., Alcón, P. & Montoya, G. Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. *Nature* **546**, 559–563 (2017).

16. Strohkendl, I., Saifuddin, F. A., Rybarski, J. R., Finkelstein, I. J. & Russell, R. Kinetic basis for DNA target specificity of CRISPR-Cas12a. *Mol. Cell* **71**, 816–824 (2018).

17. Foster, K., Grüschow, S., Bailey, S., White, M. F. & Terns, M. P. Regulation of the RNA and DNA nuclease activities required for *Pyrococcus furiosus* type III-B CRISPR–Cas immunity. *Nucleic Acids Res.* **48**, 4418–4434 (2020).

18. Wang, B. et al. Structural basis for self-cleavage prevention by tag:anti-tag pairing complementarity in type VI Cas13 CRISPR systems. *Mol. Cell* **81**, 1100–1115 (2021).

19. Meeske, A. J. & Marraffini, L. A. RNA guide complementarity prevents self-targeting in type VI CRISPR systems. *Mol. Cell* **71**, 791–801 (2018).

20. Swarts, D. C., van der Oost, J. & Jinek, M. Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR-Cas12a. *Mol. Cell* **66**, 221–233 (2017).

21. Pausch, P. et al. DNA interference states of the hypercompact CRISPR–CasΦ effector. *Nat. Struct. Mol. Biol.* **28**, 652–661 (2021).

22. Yang, H., Gao, P., Rajashankar, K. R. & Patel, D. J. PAM-dependent target DNA recognition and cleavage by C2c1 CRISPR-Cas endonuclease. *Cell* **167**, 1814–1828 (2016).

23. Zhang, H., Li, Z., Xiao, R. & Chang, L. Mechanisms for target recognition and cleavage by the Cas12i RNA-guided endonuclease.

Nat. Struct. Mol. Biol. **27**, 1069–1076 (2020).

24. Stella, S. et al. Conformational activation promotes CRISPR-Cas12a catalysis and resetting of the endonuclease activity. *Cell* **175**, 1856–1871 (2018).

25. Swarts, D. C. & Jinek, M. Mechanistic insights into the cis- and trans-acting DNase activities of Cas12a. *Mol. Cell* **73**, 589–600 (2019).

26. Swarts, D. C. Making the cut(s): how Cas12a cleaves target and non-target DNA. *Biochem. Soc. Trans.* **47**, 1499–1510 (2019).

27. Knott, G. J. et al. Structural basis for AcrVA4 inhibition of specific CRISPR-Cas12a. *Elife* **8**, e49110 (2019).

28. Gao, P., Yang, H., Rajashankar, K. R., Huang, Z. & Patel, D. J. Type v CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. *Cell Res.* **26**, 901–913 (2016).

29. Marino, N. D., Pinilla-redondo, R. & Bondy-denomy, J. CRISPR-Cas12a targeting of ssDNA plays no detectable role in immunity. (2022).

30. Meeske, A. J., Nakandakari-Higa, S. & Marraffini, L. A. Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241–245 (2019).

31. Górecka, K. M. et al. RuvC uses dynamic probing of the Holliday junction to achieve sequence specificity and efficient resolution. *Nat. Commun.* **10**, 4102 (2019).

32. Bravo, J. P. K. et al. Structural basis for mismatch surveillance by CRISPR/Cas9. *Nature* **603**, 343–347 (2022).

33. Huang, X. et al. Structural basis for two metal-ion catalysis of DNA cleavage by Cas12i2. *Nat. Commun.* **11**, 5241 (2020).

34. Gong, S., Yu, H. H., Johnson, K. A. & Taylor, D. W. DNA unwinding is the primary determinant of CRISPR-Cas9 activity. *Cell Rep.* **22**, 359–371 (2018).

35. Kurihara, N. et al. Cryo-EM structure of Cas12c2 unique pre-crRNA processing mechanism Structure of the type V-C CRISPR-Cas effector enzyme. *Mol. Cell* **82** (2022).

36. Liu, J. J. et al. CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* **566**, 218–223 (2019).

37. Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).

38. East-Seletsky, A. et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**, 270–273 (2016).

39. Liu, L. et al. The molecular architecture for RNA-guided RNA cleavage by Cas13a. *Cell* **170**, 714–726 (2017).

40. Williams, M. C. et al. Phage genome cleavage enables resuscitation from Cas13-induced bacterial dormancy. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.05.498905.

41. Sambrook, J. & Russell, D. W. *Molecular Cloning: a Laboratory Manual* (2001).

42. Harris, C. J., Molnar, A., Müller, S. Y. & Baulcombe, D. C. FDF-PAGE: a powerful technique revealing previously undetected small RNAs sequestered by complementary transcripts. *Nucleic Acids Res.* **43**, 7590–7599 (2015).

43. Punjani, A. Real-time cryo-EM structure determination. *Microsc. Microanal.* **27**, 1156–1157 (2021).

44. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: algorithms for rapid unsupervised cryo-EM structure

determination. *Nat. Methods* **14**, 290–296 (2017).

45. Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* **17**, 1214–1221 (2020).

46. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

47. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).

48. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

49. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D* **74**, 519–530 (2018).

50. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).

51. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

52. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

53. Zi Tan, Y. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).

Acknowledgements

We thank members of the laboratories of R.N.J. and D.W.T. for helpful discussions, A. Brilot for expert cryo-EM technical assistance and O. Dmytrenko for initial guidance on the in vivo plasmid interference assays. Data were collected at the Sauer Structural Biology Laboratory at the

University of Texas at Austin. This work was supported in part by National Institute of General Medical Sciences of the National Institutes of Health R35GM138348 (to D.W.T.) and R35GM138080 (to R.N.J.), an ERC Consolidator grant (865973 to C.L.B.), the Federal Agency for Disruptive Innovation (to C.L.B.), Welch Foundation Research Grant F-1938 (to D.W.T.) and a Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation Medical Research Grant (to D.W.T.). D.W.T is a CPRIT Scholar supported by the Cancer Prevention and Research Institute of Texas (RR160088) and an American Cancer Society Research Scholar supported by the American Cancer Society (RSG-21-050-01-DMC).

Author information

Author notes

1. These authors contributed equally: Jack P. K. Bravo, Thomson Hallmark

Authors and Affiliations

1. Department of Molecular Biosciences, University of Texas at Austin, Austin, TX, USA

   Jack P. K. Bravo & David W. Taylor

2. Department of Chemistry and Biochemistry, Utah State University, Logan, UT, USA

   Thomson Hallmark, Bronson Naegle & Ryan N. Jackson

3. Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Centre for Infection Research (HZI), Würzburg, Germany

   Chase L. Beisel

4. Medical Faculty, University of Würzburg, Würzburg, Germany

   Chase L. Beisel

5. Interdisciplinary Life Sciences Graduate Program, University of Texas at Austin, Austin, TX, USA

David W. Taylor

6. Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, TX, USA

David W. Taylor

7. LIVESTRONG Cancer Institutes, Dell Medical School, Austin, TX, USA

David W. Taylor

Contributions

J.P.K.B. carried out cryo-EM structure determination, model building and structural analysis. T.H. and B.N. purified and reconstituted the enzyme complexes. T.H. carried out biochemical and in vivo experiments. C.L.B. aided with in vivo experiments. J.P.K.B., T.H., R.N.J. and D.W.T. analysed and interpreted the data and wrote the manuscript. R.N.J. and D.W.T. supervised the study and obtained financial support for the work.

Corresponding authors

Correspondence to Ryan N. Jackson or David W. Taylor.

Ethics declarations

Competing interests

J.P.K.B, T.H., R.N.J. and D.W.T. are inventors on a patent application based on this research titled 'Compositions and methods related to modified Cas12a2 molecules' filed by the Board of Regents, The University of Texas System. The US Patent and Trademark Office (USPTO) has assigned US

application no. 63/349,225 to this application, and the filing date of 6 June 2022.

Peer review

Peer review information

Nature thanks Raymond Staals and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Cryo-EM data analysis.

a, Gold standard Fourier Shell Correlation (FSC) curves for Cas12a2 binary complex. **b**, Euler plot for Cas12a2 binary complex. **c**, Directional FSC[53] for Cas12a2 binary complex. **d**, Cas12a2 binary complex colored by local resolution. Left – surface, right, cut-through. **e**, Map-to-model FSC for Cas12a2 binary complex. **f**, Gold standard FSC for Cas12a2 ternary complex. **g**, Euler plot for Cas12a2 ternary complex. **h**, Directional FSC for Cas12a2 ternary complex. **i**, Cas12a2 ternary complex colored by local resolution. Left – surface, right, cut-through. **j**, Map-to-model FSC for Cas12a2 ternary complex. **k**, Gold standard FSC for Cas12a2 quaternary complex. **l**, Euler plot for Cas12a2 quaternary complex. **m**, Directional FSC for Cas12a2 quaternary complex. **n**, Cas12a2 quaternary complex colored by local resolution. Left – surface, right, cut-through. **o**, Map-to-model FSC for Cas12a2 quaternary complex.

Extended Data Fig. 2 Comparison of Cas12a2 and Cas12a.

a - c, Cas12a2 binary complex (**a**), active quaternary complex (**b**) colored by structural domain (**c**). **d – f**, Cas12a binary complex (**d**), active ternary complex (**e**) colored by structural domain (**f**). **g**, Overlay of 5′ crRNA handle, WED, RuvC and PI domains of Cas12a2 (colored) and Cas12a (transparent grey).

Extended Data Fig. 3 Structural basis for Anti-CRISPR evasion by Cas12a2.

Top – Cas12a binary complex (PDB 5NG6, left) and Cas12a2 (right) with AcrVA1 (PDB 6NMD, green). AcrVA1 interacts with the 5′ crRNA seed in Cas12a, which is not present in Cas12a2. There are also significant clashes between AcrVA1 and the REC lobe of Cas12a2. Middle – Cas12a (left) and Cas12a2 (right) binary complex with a monomer of AcrVA4 (PDB 6NMA, orange). There are severe clashes between the Cas12a2 Insertion domain (dark cyan) and AcrVA4. Clashes are shown as magenta lines.

Extended Data Fig. 4 Mechanism of protospacer-flanking sequence (PFS) recognition by Cas12a2.

a, Schematic representation of PFS (GAAAG) located at the 3′ end of the RNA target RNA. **b**, Positions of the PFS within the Cas12a2 active ternary complex. **c**, Zoom-in of interactions between Cas12a2 and the first four bases of PFS. **d**, Activation of Cas12a2 or Cas12a2 ΔPI cleavage by truncated target RNAs. Representative of three independent experiments with similar results. **e**, Circular Dichroism (CD) spectroscopy of Cas12a2 mutants, including ΔPI truncation. All mutants are properly folded. For gel source data, see Supplementary Fig. 1.

Source Data

Extended Data Fig. 5 Conformational changes in Insertion domain upon target RNA binding.

a, Cas12a2 autoinhibited binary complex (left) and active ternary complex (right), colored as in Fig. 1a. Insertion domain is dark cyan. **b**, conformational changes of the C-terminal half of insertion domain from

binary (grey) to ternary (dark cyan) complexes. N-terminal half remains static, and interacts with the crRNA 5′ handle.

Extended Data Fig. 6 Collateral ssDNA in other Cas12 structures.

Top, Cas12a2 ternary complex (this study), Cas12b (PDB 5U31) and Cas12i (6W5C). Arrow denotes Cas12a2 active site. Boxes show position of bound collateral ssDNA in Cas12b and Cas12i. **Bottom**, Zoom-in of collateral ssDNA (dark cyan) buried in RuvC active site of Cas12b and Cas12i.

Extended Data Fig. 7 Representative cryo-EM densities.

a, target RNA and crRNA from ternary complex map, showing 5′ crRNA handle and 3′ target RNA PFS. **b**, 5′ handle and 7-nt pre-ordered seed from Cas12a2 binary complex. **c**, Cleaved strand pair of aromatic clamps. **d**, non-cleaved strand (NCS) aromatic clamps. Density for the unwound bases of the NCS is diffuse due to flexibility, but the nucleobase held within the aromatic clamp is well-resolved. **e**, RuvC active site, including density for scissile phosphate, two Mg^{2+} ions and a putative activating water.

Extended Data Fig. 8 Effect of clamp residue mutations on supercoiled DNA cleavage *in vitro*.

Rate of supercoiled DNA cleavage by wild-type Cas12a2 and the four aromatic clamp mutants described in Fig. 3. WT Cas12a2 completely cleaves supercoiled plasmid within 10 min, while all four mutants significantly reduce complete cleavage. Y465A retains an ability (albeit much slower than WT Cas12a2) to nick and/or linearize plasmid, but fails to fully degrade plasmid DNA. **a**, Quantitation of fraction of supercoiled pUC19 cleaved over time. Significance between WT and mutant SuCas12a2 was determined by two-sided Student T-test. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Experiments were performed in triplicate, and error bars correspond to the mean and standard error. **b**. Representative agarose gels of supercoiled pUC19 cleavage by WT Cas12a2 or aromatic clamp mutants over time. Gels are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.

Source Data

Extended Data Fig. 9 *In vivo* interference assays.

Transformation efficiency of Target (T) and Non-Target (NT) plasmids into electrocompetent *E. coli* containing WT or mutant Cas12a2+3xcrRNA immune system plasmids shown by spot assay. Transformations were plated as 10x serial dilutions between 10^1 and 10^6 and the number of colonies in the highest countable spot were used to determine the total number colony forming units (cfu) per 50 ng plasmid. Transformation fold reduction was calculated as $-Log_{10}(T_{cfu}/NT_{cfu})$. Active immune systems present a ≥100-fold difference between T and NT transformation efficiency. All transformations were completed to at least two biological replicates. For transformation plates source data, see Supplementary Fig. 1.

Extended Data Fig. 10 Effect of pre-incubating activating RNA with Cas12a2 on collateral ssDNA and ssRNA cleavage.

Cas12a2 cleavage activity after either No or 2 h incubation of Cas12a2:crRNA complex with 2x Target RNA. **A**. Non-target cleavage of DNAse Alert reporter (IDT) **B**. Non-target cleavage of RNase Alert reporter (IDT). Each experiment was performed in triplicate. For gel source data, see Supplementary Fig. 1.

Source Data

Extended Data Table 1 Cryo-EM data collection and model validation statistics

Supplementary information

Supplementary Information

This file contains Supplementary Fig. 1 and Table 1 and the legend to Supplementary Video 1.

Reporting Summary

Peer Review File

Supplementary Video 1

Source data

Source Data Fig. 3 and Source Data Extended Data Figs. 4, 8 and 10

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Bravo, J.P.K., Hallmark, T., Naegle, B. *et al.* RNA targeting unleashes indiscriminate nuclease activity of CRISPR–Cas12a2. *Nature* **613**, 582–587 (2023). https://doi.org/10.1038/s41586-022-05560-w

- Received: 13 June 2022
- Accepted: 11 November 2022
- Published: 04 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05560-w

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

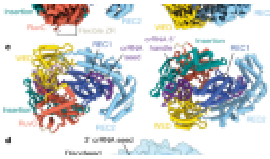
Provided by the Springer Nature SharedIt content-sharing initiative

This article is cited by

- **Cas12a2 elicits abortive infection through RNA-triggered destruction of dsDNA**
  - Oleg Dmytrenko
  - Gina C. Neumann
  - Chase L. Beisel

  Nature (2023)

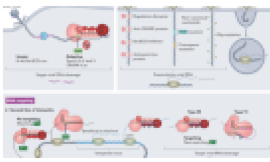
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05560-w



RNA targeting unleashes indiscriminate nuclease activity of CRISPR–Cas12a2

04 January 2023

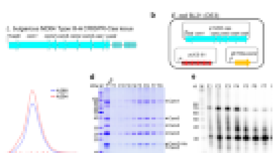
Jack P. K. Bravo, Thomson Hallmark, … David W. Taylor



RNA-targeting CRISPR–Cas systems

28 September 2022

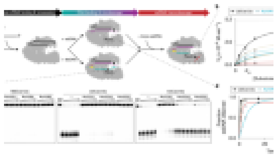
Sam P. B. van Beljouw, Jasper Sanders, … Stan J. J. Brouns



Characterization of a novel type III CRISPR-Cas effector provides new insights into the allosteric activation and suppression of the Cas10 DNase

12 May 2020

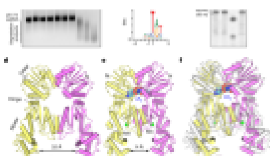
Jinzhong Lin, Mingxia Feng, … Qunxin She



Broad-spectrum enzymatic inhibition of CRISPR-Cas12a

01 April 2019

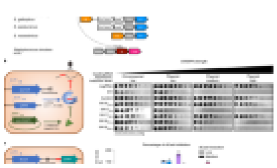
Gavin J. Knott, Brittney W. Thornton, … Jennifer A. Doudna



The Card1 nuclease provides defence during type III CRISPR immunity

18 January 2021

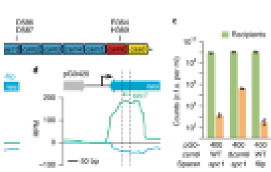
Jakob T. Rostøl, Wei Xie, … Luciano A. Marraffini



Broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer

26 March 2020

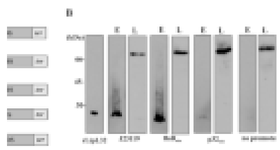
Caroline Mahendra, Kathleen A. Christie, … Joseph Bondy-Denomy



Non-specific degradation of transcripts promotes plasmid clearance during type III-A CRISPR–Cas immunity

28 January 2019

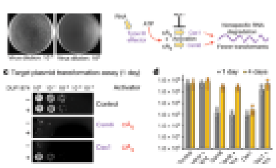
Jakob T. Rostøl & Luciano A. Marraffini



Gene silencing based on RNA-guided catalytically inactive Cas9 (dCas9): a new tool for genetic engineering in Leptospira

12 February 2019

L. G. V. Fernandes, L. P. Guaman, … A. L. T. O. Nascimento



An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity

15 January 2020

Januka S. Athukoralage, Stephen A. McMahon, … Malcolm F. White

- Article
- Open Access
- Published: 04 January 2023

Cas12a2 elicits abortive infection through RNA-triggered destruction of dsDNA

- Oleg Dmytrenko,
- Gina C. Neumann,
- Thomson Hallmark,
- Dylan J. Keiser,
- Valerie M. Crowley,
- Elena Vialetto,
- Ioannis Mougiakos,
- Katharina G. Wandera,
- Hannah Domgaard,
- Johannes Weber,
- Thomas Gaudin,
- Josie Metcalf,
- Benjamin N. Gray,
- Matthew B. Begemann ,
- Ryan N. Jackson &
- …
- Chase L. Beisel

Show authors

Nature **volume 613**, pages 588–594 (2023)

- 11k Accesses
- 1 Citations
- 416 Altmetric
- Metrics details

Abstract

Bacterial abortive-infection systems limit the spread of foreign invaders by shutting down or killing infected cells before the invaders can replicate[1,2]. Several RNA-targeting CRISPR–Cas systems (that is, types III and VI) cause abortive-infection phenotypes by activating indiscriminate nucleases[3,4,5]. However, a CRISPR-mediated abortive mechanism that leverages indiscriminate DNase activity of an RNA-guided single-effector nuclease has yet to be observed. Here we report that RNA targeting by the type V single-effector nuclease Cas12a2 drives abortive infection through non-specific cleavage of double-stranded DNA (dsDNA). After recognizing an RNA target with an activating protospacer-flanking sequence, Cas12a2 efficiently degrades single-stranded RNA (ssRNA), single-stranded DNA (ssDNA) and dsDNA. Within cells, the activation of Cas12a2 induces an SOS DNA-damage response and impairs growth, preventing the dissemination of the invader. Finally, we harnessed the collateral activity of Cas12a2 for direct RNA detection, demonstrating that Cas12a2 can be repurposed as an RNA-guided RNA-targeting tool. These findings expand the known defensive abilities of CRISPR–Cas systems and create additional opportunities for CRISPR technologies.

Main

All domains of life use defence strategies that cause cells to enter dormancy or die to limit the spread of infectious agents[1]. In bacteria and archaea, this strategy is called abortive infection, and it is used by a vast variety of bacterial defence systems[1,2]. Recently, it was shown that CRISPR RNA (crRNA)-guided adaptive immune systems that target RNA cause abortive-infection phenotypes[3,4,5,6]. Type VI systems non-specifically degrade RNA, whereby the Cas13 single-effector nuclease acts as both a crRNA-guided effector and indiscriminate RNase[3,7,8]. In type III systems, target RNA binding triggers the production of cyclic oligoadenylate secondary messengers that in turn activate indiscriminate accessory RNases and ssDNases that can drive abortive infection[4,5,9,10,11]. Moreover, it has been proposed that abortive infection is mediated by indiscriminate dsDNases (such as NucC) activated through type III secondary messengers[12,13] or by

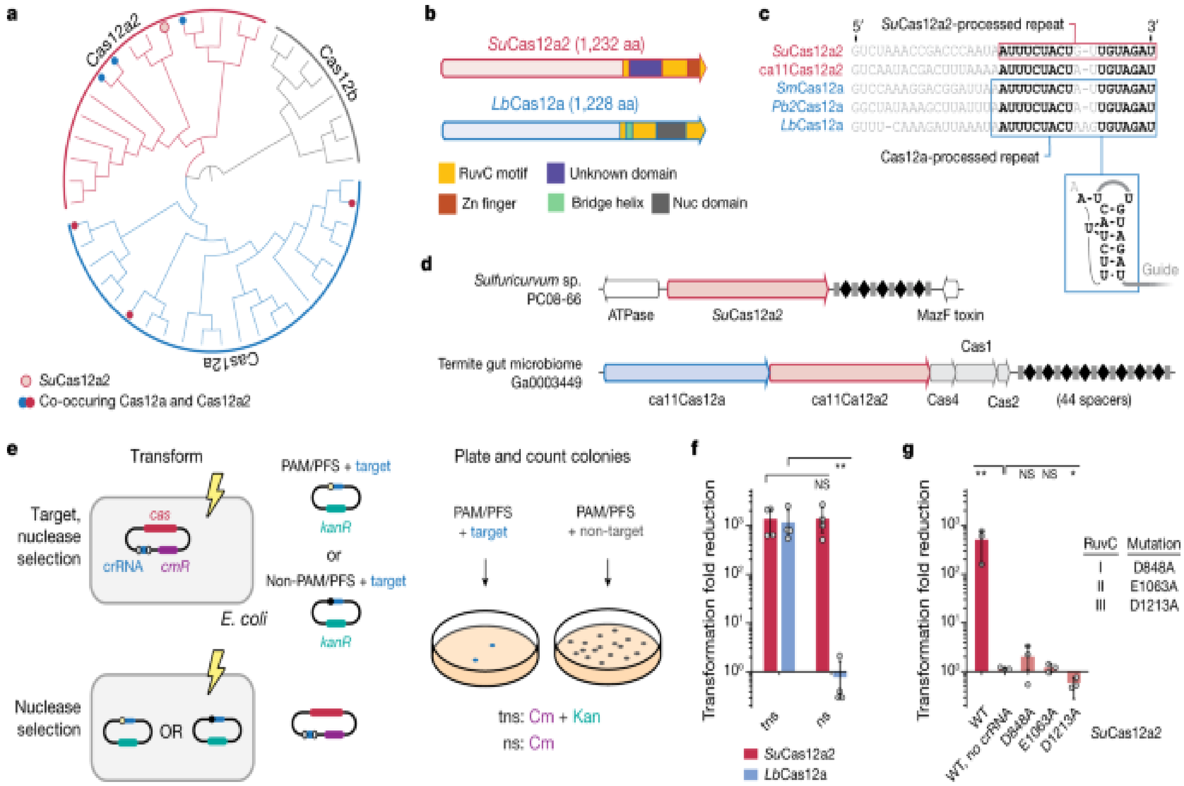
indiscriminate ssDNase activity from type V Cas12a single-effector nucleases[14]. However, type III CRISPR-mediated dsDNase activity has yet to be examined in vivo, and the ssDNase activity of Cas12a was recently shown to not cause abortive infection[15].

Here we report that Cas12a2, a type V single-effector CRISPR-associated (Cas) nuclease, induces an abortive-infection phenotype when challenged with plasmids that are complementary to crRNA guides. Biochemical assays using recombinant protein revealed that Cas12a2 recognizes RNA targets, unleashing non-specific dsDNA-, ssDNA- and ssRNA-nuclease activities distinct from those of other single-subunit RNA-targeting (such as Cas13a) and dsDNA-targeting (such as Cas12a) Cas nucleases[8,16,17]. Furthermore, we show that the Cas12a2 non-specific nuclease activities damage bacterial DNA, triggering the SOS response and impairing cell growth. Collectively these results suggest that the dsDNase activity of Cas12a2 is instrumental in triggering the abortive-infection phenotype. As a proof-of-principle demonstration, we show that Cas12a2 can detect RNA at a sensitivity that is comparable to that of the RNA-targeting Cas13a nuclease at various temperatures.

Cas12a2 induces abortive infection

Cas12a2 comprises a group of type V effector nucleases that are related to Cas12a[16], with Cas12a2 orthologues previously being classified as Cas12a variants[18]. Our analyses similarly place them in a monophyletic clade that shares the last common ancestor with Cas12a nucleases (Fig. 1a and Extended Data Fig. 1). Further analysis revealed that CRISPR–Cas12a2 and CRISPR–Cas12a systems feature CRISPR repeats with a conserved 3′ end, and the nucleases possess homologous RuvC endonuclease domains and a similar predicted secondary structure in the N termini (Fig. 1b and Supplementary Fig. 2). Despite the conserved RuvC-like domains and N termini, Cas12a2 is distinguished from Cas12a by the presence of a large domain of unknown function located in place of the Cas12a bridge helix as well as a zinc-finger domain in place of the Cas12a Nuc domain (Fig. 1b and Supplementary Fig. 2). Considering their original classification combined with our phylogenetic analyses as well as recent structural results[19], we named these distinct type V nucleases Cas12a2.

Fig. 1: Cas12a2 nucleases form a distinct clade within type V Cas12 nucleases.



a, Maximum-likelihood phylogeny of identified Cas12a2 nucleases with Cas12a and Cas12b nucleases. The detailed phylogeny is shown in Extended Data Fig. 1. Systems with co-occurring Cas12a2 and Cas12a are indicated by filled red and blue circles. *Su*Cas12a2 is indicated by an unfilled red circle. **b**, The domain architecture of *Su*Cas12a2 in comparison to *Lb*Cas12a. aa, amino acids. **c**, Aligned direct repeats associated with representative Cas12a2 and Cas12a nucleases. The bold nucleotides indicate conserved positions within the processed repeats for both nucleases. The predicted pseudoknot structure of the Cas12a repeat is shown below. The loop of the hairpin (grey) is variable. Pre-crRNA processing by *Su*Cas12a2 is shown in Extended Data Fig. 3. **d**, Gene organization of CRISPR–Cas systems within representative genomic loci encoding Cas12a2. Examples of systems encoding Cas12a2 as the sole Cas nuclease and those also encoding Cas12a are shown. **e**, Diagram of the traditional (top; target nuclease selection (tns)) and modified (bottom; nuclease selection (ns)) plasmid interference assay. Cm, chloramphenicol; Kan, kanamycin. **f**, The reduction in plasmid transformation for *Su*Cas12a2 and *Lb*Ca12a2 under target plasmid and nuclease plasmid selection. **g**, The reduction in plasmid transformation of

*Su*Cas12a2 RuvC mutants under target plasmid and nuclease plasmid selection. For **f** and **g**, data are mean ± s.d. of at least three independent experiments started from separate colonies. *P* values were calculated using one-tailed Welch's *t*-tests; NS, $P > 0.05$; $^{*}P < 0.05$, $^{**}P < 0.005$.

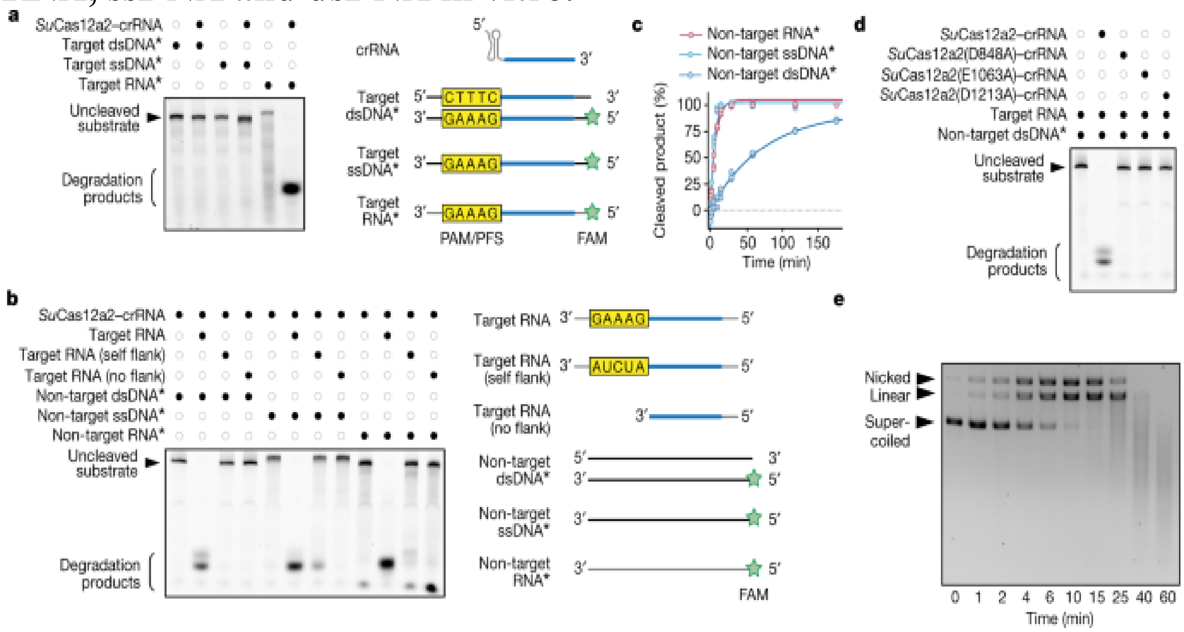
Notably, some CRISPR–Cas systems contain both *cas12a2* and *cas12a* genes in tandem next to a shared CRISPR array (Fig. 1c). From this observation and the conservation of CRISPR repeats from systems with either of the nucleases (Fig. 1d and Supplementary Fig. 3), we hypothesized that both proteins bind to and process similar crRNA guides. However, as the proteins diverge in other domains, we further hypothesized that Cas12a2 performs a defence function that is distinct from the dsDNA-targeting activity of Cas12a[16].

To test these hypotheses, we encoded the *cas12a2* gene from the sulfur-oxidizing epsilonproteobacterium *Sulfuricurvum* sp. PC08-66 (*Su*Cas12a2) along with a CRISPR array into an expression plasmid, which we introduced into *Escherichia coli* cells. We next performed a traditional plasmid interference assay that depletes cells by selecting for the plasmid containing the nuclease and crRNA as well as a target plasmid (Fig. 1e). This assay detects broad immune system activity but cannot distinguish between defence activities that only deplete the target from those that activate abortive-infection phenotypes. To test whether Cas12a2 uses an abortive-infection mechanism, we modified the assay by selecting only for the nuclease plasmid (Fig. 1e). Consistent with our hypothesis that Cas12a2 functions differently compared with Cas12a, Cas12a2 depleted cells in both the traditional (about 1,900-fold reduction) and modified (around 1,300-fold reduction) plasmid interference assays, whereas Cas12a from *Lachnospiraceae bacterium* (*Lb*Cas12a) depleted cells only in the traditional assay (Fig. 1f). Similar trends were observed with different targets cloned into the same plasmid location, different Cas12a2 homologues and when comparing *Su*Cas12a2 with the Cas12a homologue from *Prevotella bryantii B14* (*Pb*2Cas12a) (Extended Data Fig. 2a,b). Moreover, mutating predicted active residues within any of the three RuvC motifs in *Su*Cas12a2 impaired immune function (Fig. 1g). Collectively, these results indicate that Cas12a2 relies on a RuvC-nuclease domain and induces abortive infection through a mechanism that is distinct from that of Cas12a.

Cas12a2 targets RNA and degrades dsDNA

CRISPR systems that cause abortive-infection phenotypes (such as types III and VI) rely on indiscriminate nucleases activated by RNA targeting[3,4,5]. To determine whether Cas12a2 uses a similar mechanism, we recombinantly expressed and purified *Su*Cas12a2 and tested its enzymatic activities in vitro (Fig. 2 and Supplementary Fig. 4). However, before examining the nucleic-acid-targeting activities, we needed to determine how Cas12a2 crRNAs are processed.

Fig. 2: RNA target recognition by *Su*Cas12a2 triggers degradation of ssRNA, ssDNA and dsDNA in vitro.



a, Direct targeting of different FAM-labelled nucleic-acid substrates by a purified *Su*Cas12a2–crRNA complex. **b**, Collateral cleavage of FAM-labelled non-target nucleic-acid substrates by the *Su*Cas12a2–crRNA complex with different target RNA substrates after 1 h. Target RNA, a non-self flanking sequence at the 3′ end; self flank, a flanking sequence mutated to the reverse complement of the crRNA repeat tag; no flank, only the reverse complement of the crRNA guide. For **a** and **b**, diagrams of target and non-target nucleic acids are shown on the right. **c**, Time-course analysis of RNA-triggered collateral cleavage of labelled non-target RNA, ssDNA or dsDNA. Representative gel images are provided in Extended Data Fig. 4c. Note that dsDNA contains twice as much ssDNA substrate as the RNA and

ssDNA, but the same concentration of labelled strands. **d**, The effect of mutating each of the three RuvC motifs on RNA-triggered collateral cleavage of dsDNA. **e**, Time-course analysis of RNA-triggered collateral cleavage of non-target plasmid DNA. Plasmid DNA was visualized using ethidium bromide. For **a**–**d**, the asterisks indicate a FAM-labelled substrate, and the diagrams on the right indicate the substrates. All of the results are representative of three independent experiments. Gel source data are provided in Supplementary Fig. 1.

The CRISPR repeats of the Cas12a and Cas12a2 systems are highly conserved at the 3′ end (Fig. 1d and Supplementary Fig. 3), and sequence alignments predict that Cas12a2 shares secondary structure in the region of the Cas12a pre-crRNA-processing active site[18,20] (Supplementary Fig. 5). Consistent with this prediction, RNA-sequencing analysis of pre-crRNAs processed by *Su*Cas12a2 in vitro revealed that processing occurs one nucleotide downstream of the position cleaved by Cas12a (Extended Data Fig. 3a,b). The 3′ end of the spacer also underwent trimming in vivo to form an approximately 24-nucleotide guide (Extended Data Fig. 3b,c), possibly through host ribonucleases as observed for Cas9 crRNAs[21]. Mutating basic amino acids (Lys784 and Arg785) located in the predicted RNA-processing active site abolished activity[22] (Extended Data Fig. 3d). Furthermore, plasmid interference assays revealed that Cas12a and Cas12a2 can interchange guides without impairing immunity (Extended Data Fig. 2c). Thus, the Cas12a2 nuclease processes its own crRNA guides like other type V effector nucleases[20,23] and can share crRNAs with Cas12a.

To determine the nucleic-acid target preference of crRNA-guided Cas12a2, complementary ssRNA, ssDNA and dsDNA substrates containing an A/T-rich flanking sequence (paralleling Cas12a substrates)[16,22] were fluorescently labelled with a FAM molecule and combined with crRNA-guided Cas12a2 (Fig. 2a). Similar to CRISPR–Cas systems that cause abortive infection—yet in contrast to the dsDNA-targeting Cas12a—Cas12a2 is activated only in the presence of complementary RNA targets. The potency of plasmid interference with *Su*Cas12a2 (Fig. 1f) was notable given the lack of a defined promoter upstream of the target in this construct. However, we attribute the interference to spurious transcription of the encoding plasmid for two reasons: introducing an upstream terminator

significantly reduced plasmid interference in *E. coli* (Extended Data Fig. 2d,e), and an upstream promoter was required to detect collateral activity in a cell-free transcription–translation assay[24] (Extended Data Fig. 2f,g).

As other Cas abortive-infection mechanisms rely on collateral indiscriminate RNase activity, we examined whether specific RNA targeting by Cas12a2 induces indiscriminate nuclease activity. We found that *Su*Cas12a2 robustly and indiscriminately degraded FAM-labelled ssRNA, ssDNA and dsDNA substrates bearing no complementarity to the crRNA guide. By contrast, other Cas nucleases indiscriminately degrade only ssRNA (Cas13a)[8] or ssRNA and ssDNA (Cas12g)[25] after RNA targeting, or only ssDNA after dsDNA targeting (Cas12a)[14] (Fig. 2b and Extended Data Fig. 4a). Of the three collateral substrates, ssRNA and ssDNA appear to be more efficiently cleaved than dsDNA by Cas12a2 (Fig. 2c and Extended Data Fig. 4b). However, this difference could be explained by the presence of twice as many DNA strands in dsDNA substrates compared with ssDNA substrates for the same amount of nuclease. Also, similar to Cas13a[8], complementary ssDNA and dsDNA do not activate any Cas12a2 non-specific nuclease activity (Extended Data Fig. 4a), and dsRNA is not a primary substrate of collateral cleavage (Extended Data Fig. 4c).

To examine whether Cas12a2 activity is reliant on detecting a 'non-self' signal adjacent to the target (called a protospacer-flanking sequence (PFS))[8], we performed in vitro cleavage assays in which target RNA sequences were flanked on the 3′ side with a 'self' sequence complementary to the crRNA repeat (5′-AUCUA-3′), the non-self PFS used in our in vivo assay (5′-GAAAG-3′) or a 'flankless' RNA complementary to the guide region of the crRNA, but containing no PFS (Fig. 2b). Notably, only the RNA target containing the non-self PFS activated collateral nuclease activity, demonstrating that specific nucleotides at the 3′ end of the RNA target must be present to activate the collateral activity of Cas12a2. Moreover, introducing disruptive mutations to any of the three RuvC motifs or conserved cysteine residues within the putative zinc-finger domain abolished all non-specific cleavage (Fig. 2d and Extended Data Fig. 4d), consistent with our in vivo plasmid interference results (Fig. 1g).

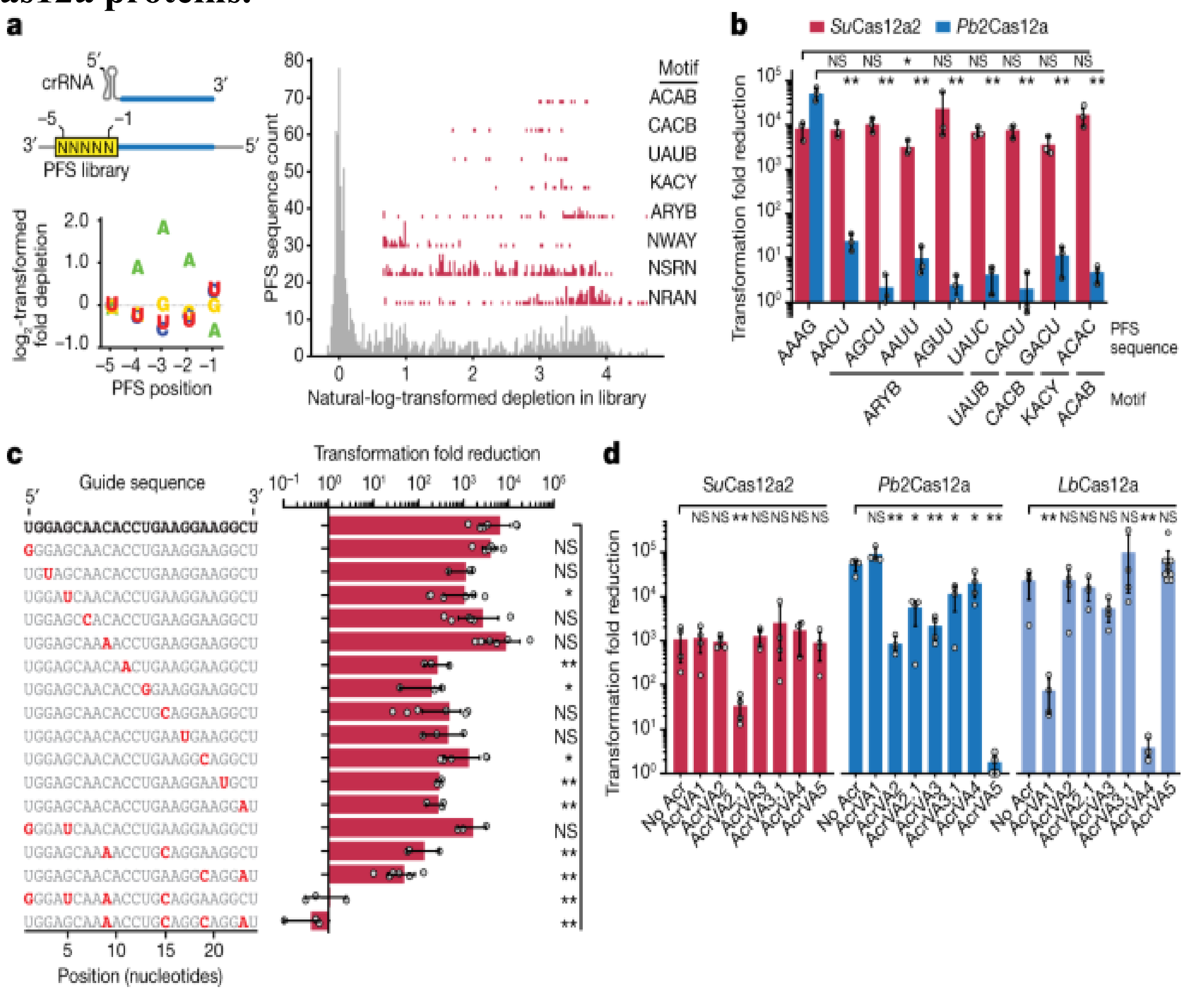
Our biochemical assays demonstrated that Cas12a2 could quickly remove a FAM label from linear dsDNA substrates, but it was unclear whether Cas12a2 degrades DNA lacking available 5′ or 3′ ends. We therefore challenged crRNA-guided Cas12a2 with an RNA target and a supercoiled pUC19 plasmid. Importantly, pUC19 does not contain any sequence complementary to the Cas12a2 crRNA guide. We observed that *Su*Cas12a2 rapidly nicked, linearized and degraded pUC19 DNA (Fig. 2e), but only in the presence of a cognate target and PFS and with an intact RuvC domain (Extended Data Fig. 4e). This rapid destruction of the supercoiled plasmid contrasts with the slow and incomplete linearization of plasmid DNA by Cas12a nucleases[26]. These data suggest a mechanism in which activated *Su*Cas12a2 is able to robustly hydrolyse the phosphodiester backbone of non-specific DNA regardless of whether it is supercoiled, nicked or linear. A comparison with Cas12a (dsDNA targeting with collateral ssDNase), Cas13a (ssRNA targeting with collateral ssRNase) and Cas13g (ssRNA targeting with collateral ssRNase and ssDNase) demonstrated that the RNA-targeting ssRNase, ssDNase and dsDNase are unique to *Su*Cas12a2 (Extended Data Fig. 4a). Collectively, these in vitro results reveal that crRNAs guide *Su*Cas12a2 to RNA targets, activating RuvC-dependent cleavage of ssRNA, ssDNA and dsDNA. These activities, in part or in total, may underlie the abortive-infection phenotype.

Cas12a2 exhibits targeting flexibility

Although our in vitro data indicated an underlying mechanism for the Cas12a2 abortive-infection phenotype, we wanted to understand the targeting limitations of these distinct enzymes. In particular, we investigated the stringency of non-self PFS sequence recognition and penalties for mismatches between the crRNA and target. We therefore challenged *Su*Cas12a2 with a library of plasmids encoding all possible 1,024 flanking sequences at the 3′ end of the RNA target to the −5 position (Fig. 3a and Extended Data Fig. 5a,b). We found that *Su*Cas12a2 depleted approximately half of all of the sequences in the library, suggesting a PFS-recognition mechanism that is more stringent than that of Cas13 but still more promiscuous than those of most DNA-targeting systems[8,27]. The depleted sequences were generally A rich, consistent with a 5′-GAAAG-3′ PFS, but could not be fully captured by a single consensus motif (Fig. 3a and

Extended Data Fig. 5c). We further validated individual depleted sequences, including representatives within five unique motifs recognized by *Su*Cas12a2 but not by *Pb*2Cas12a—a nuclease that is known for flexible PAM recognition[24] (Fig. 3b and Extended Data Fig. 5d). Consistent with its function as an RNA-targeting nuclease, the recognized sequences were broad but did not follow the expected profile if Cas12a2 is principally evaluating tag–anti-tag complementarity. These results further support a mechanism in which PFS recognition by *Su*Cas12a2 operates similar to type III systems that require recognition of a PFS or RNA PAM to activate[28,29,30,31] and distinct from the evaluation of tag–anti-tag complementarity used by RNA-targeting Cas13[8,32] and other type III CRISPR–Cas systems[33].

Fig. 3: *Su*Cas12a2 exhibits promiscuous targeting and resists anti-Cas12a proteins.



a, Experimentally determined PFSs and motifs recognized by *Su*Cas12a2 in *E. coli*. Motifs capturing positions −4 to −1 of the PFS are shown, and are written 3′ to 5′. B represents C, G or U; K represents G or U; R represents G or A; W represents A or U; and Y represents C or U. Results are representative of two independent screens (Extended Data Fig. 5). **b**, Validation of selected PFSs identified in the screen and permissive of targeting by *Su*Cas12a2 but not *Pb*2Cas12a. **c**, The effect of guide mismatches on plasmid targeting by *Su*Cas12a2 in *E. coli*. **d**, The extent of inhibition by known AcrVA proteins against *Su*Cas12a2. Acr proteins were confirmed to exhibit inhibitory activity against different Cas12a homologues in *E. coli* or in cell-free transcription–translation reactions (Extended Data Fig. 5). Data are mean ± s.d. of at least three independent experiments started from separate colonies. Statistical analysis was performed using one-tailed Welch's *t*-tests.

Most DNA- and RNA-targeting Cas nucleases have shown high sensitivity to mismatches within a seed region, in which a single mismatch between the crRNA guide and target disrupts binding[8,17,34,35]. Thus, to identify whether *Su*Cas12a2 relies on a seed region, we evaluated how *Su*Cas12a2 tolerates mismatches in our cell-based assay (Fig. 3c). Notably, *Su*Cas12a2 accommodated single and double mismatches across the target, with PFS-distal mutations exerting more adverse effects on plasmid targeting. To completely disrupt *Su*Cas12a2 targeting, four mismatches were needed throughout the guide (Fig. 3c) or up to 10 mismatches at the 3′ end of a 24-nucleotide guide (Extended Data Fig. 6a). The flexible PFS recognition and a tolerance for guide–target mismatches indicate that *Su*Cas12a2 exhibits promiscuous target recognition and appears to lack a canonical seed that is hypersensitive to guide–target mismatches[34,35]. However, the necessary pairing with the 3′ end of the guide is consistent with this end being pre-ordered in the structure of the crRNA–Cas12a2 binary complex[19] and possibly initiating base pairing with the target. Furthermore, the promiscuity further enabled *Su*Cas12a2 to recognize target mutations that disrupt targeting by Cas12a (Extended Data Fig. 6b).

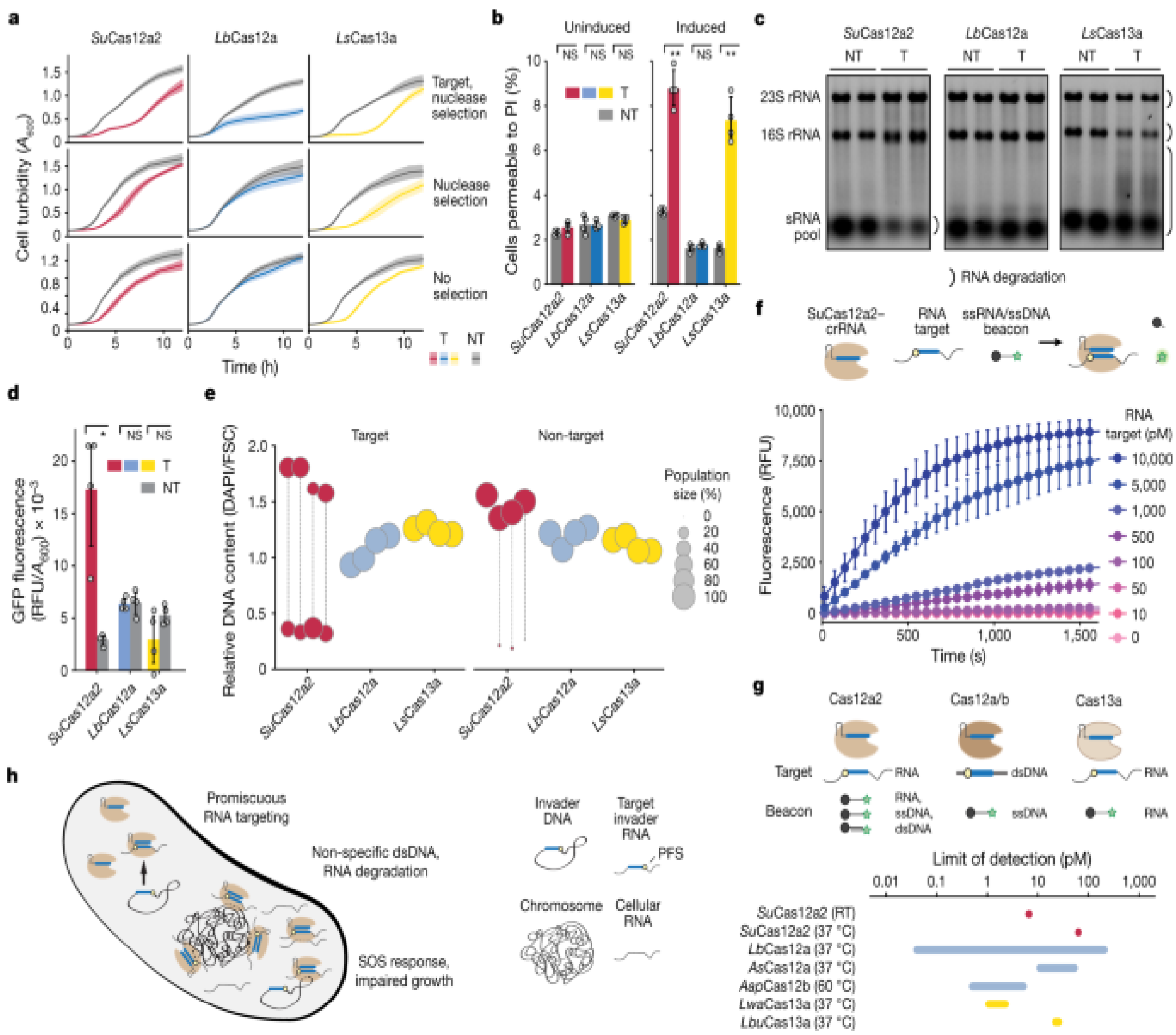
Collectively, the distinct activities of Cas12a2 compared with Cas12a suggested that tandem systems possessing both nucleases (Fig. 1d) may act cooperatively to broaden the effectiveness against foreign viruses and

plasmids. In particular, we hypothesized that the unique structural features of Cas12a2 might prevent the escape of viruses encoding anti-CRISPR proteins that block Cas12a function[36,37,38]. Consistent with this hypothesis, only one (AcrVA2.1) out of seven Cas12a anti-CRISPR proteins was able to impair Cas12a2 function, albeit only partially (Fig. 3d and Supplementary Fig. 6a,b). Notably, AcrVA5 also exhibited no inhibitory activity despite *Su*Cas12a2 possessing the conserved lysine residue in Cas12a that is chemically modified by this Acr to block PAM recognition[39] (Fig. 3d and Supplementary Fig. 6c). The limited ability of Cas12a Acr proteins to inhibit *Su*Cas12a2 further underscores the distinct properties of these nucleases and the ability of Cas12a and Cas12a2 to complement each other in immune defence.

SOS response and dormancy by Cas12a2

Although our initial results indicate that Cas12a2 causes an abortive-infection phenotype, one scenario was that triggered Cas12a2 was selectively clearing all plasmids, enabling the cells to succumb to any introduced antibiotic selection. To assess this possibility, we evaluated the growth of *E. coli* in liquid culture under different antibiotic selection conditions (including a no-antibiotic condition) after induction of *Su*Cas12a2 and *Lb*Cas12a using a targeting crRNA (Fig. 4a and Supplementary Fig. 7). *Su*Cas12a2 but not *Lb*Cas12a suppressed culture growth in the absence of plasmid selection, further supporting induction of an abortive-infection phenotype by Cas12a2.

Fig. 4: *Su*Cas12a2 causes abortive infection principally through collateral DNA degradation and can be harnessed for RNA detection.



a, *E. coli* growth arrest in the presence of *Su*Cas12a2, *Lb*Cas2a2 or *Ls*Cas13a plasmid under different targeting conditions and antibiotic regimes. A_{600}, absorbance at 600 nm. **b**, The percentage of *E. coli* cells stained with propidium iodide (PI) indicative of viability loss before targeting (uninduced) and after 4 h of targeting (induced) without antibiotic selection. The gating strategy is shown in Supplementary Fig. 7. **c**, The varying extent of RNA degradation in *E. coli* by *Su*Cas12a2, *Lb*Cas12a or *Ls*Cas13a2 2 h after induction without antibiotic selection. The results represent duplicate independent experiments. See Extended Data Fig. 7 for independent quadruplicates. The small RNA pool includes tRNAs and other small RNAs. **d**, SOS-responsive expression of GFP in *E. coli* after 4 h of plasmid targeting by *Su*Cas12a2, *Lb*Cas12a or *Ls*Cas13a without antibiotic selection. Time-course data are shown in Supplementary Fig. 9. RFU, relative fluorescence units. **e**, Relative DNA content in *E. coli* after 4 h of

targeting by *Su*Cas12a2, *Lb*Cas12a or *Ls*Cas13a without antibiotic selection. 4′,6-diamidino-2-phenylindole (DAPI) fluorescence and cell size were measured using flow cytometry. Each circle or pair of vertically aligned circles represents major subpopulations from the same biological replicate. The corresponding contour plots are shown in Extended Data Fig. 8. FSC, forward scatter. **f**, RNA-detection assay. The limit of RNA detection with Cas12a2 incubated with a ssRNA beacon at room temperature (RT). Data are mean ± s.e.m. of three independent experiments. **g**, Nucleic-acid targets and reporters as well as the unamplified limit of detection (LOD) for Cas12a2 and other Cas detectors[46]. *Aap*, *Alicyclobacillus acidiphilus*; *Lwa*, *Leptotrichia wadei*; *Lbu*, *Leptotrichia buccalis*. **h**, The proposed model for promiscuous RNA targeting and collateral degradation by *Su*Cas12a2 and its effect on the cell. For **b** and **d**, data are mean ± s.d. of four independent experiments started from separate colonies. Statistical analysis was performed using one-tailed Welch's *t*-tests. NT, non-target plasmid; T, target plasmid.

One open question is whether the abortive-infection phenotype was caused by cell dormancy or cell death. It was recently shown that, after recognizing an RNA target, Cas13a mediates widespread RNA degradation that drives cellular dormancy and suppresses phage infection[3]. We therefore introduced the representative Cas13a from *Leptotrichia shahii* (*Ls*Cas13a) into our liquid culture assay. Similar to *Su*Cas12a2, *Ls*Cas13a suppressed growth even in the absence of plasmid selection (Fig. 4a and Supplementary Fig. 7).

Our comparison to *Ls*Cas13a suggested that the growth suppression by *Su*Cas12a2 could occur through non-specific RNA cleavage, causing cell dormancy, whereas our in vitro data indicated that non-specific dsDNA cleavage could also suppress growth by causing cell death. To evaluate whether the cells containing *Su*Cas12a2 were undergoing cell death, we performed a cell-viability assay using propidium iodide for both Cas12a2 and Cas13a. We observed only a small percentage (about 10%) of cell death for both Cas12a2 and Cas13a after 4 h (Fig. 4b and Supplementary Fig. 8). Thus, although the indiscriminate nuclease activities of Cas12a2 cause some cell death, the primary result of Cas12a2 activity is better described as a cell-dormancy phenotype.

Although Cas12a2 appears to cause dormancy, it was unclear which of the several indiscriminate nuclease activities are involved. To determine whether *Su*Cas12a2 causes dormancy through RNA cleavage, total cellular RNAs were examined under targeting and non-targeting conditions. Whereas Cas13a significantly depleted both rRNAs and the small RNA pool (which includes tRNAs), Cas12a2 significantly depleted only the small RNA pool (Fig. 4c and Extended Data Fig. 7).

Given the observed differences in RNA degradation under targeting conditions, we examined whether the indiscriminate dsDNase activity of *Su*Cas12a2 was detectable in the context of the abortive-infection phenotype. We reasoned that widespread dsDNA damage caused by *Su*Cas12a2 would trigger an SOS response, impairing growth[40,41]. Consistent with this assertion, plasmid targeting using *Su*Cas12a2 significantly induced GFP expression from an SOS-responsive reporter construct[42] compared with a non-target control, whereas *Lb*Cas12a and *Ls*Cas13a negligibly induced GFP expression (Fig. 4d and Supplementary Fig. 9). Furthermore, *Su*Cas12a2-targeting cultures diverged into two subpopulations in the absence of antibiotic selection: one represented by compact cells with reduced DNA content and another represented by filamentous cells with high DNA content (Fig. 4e and Extended Data Figs. 8 and 9). Cultures expressing *Lb*Cas12a and *Ls*Cas13a did not exhibit noticeable differences in cell size and DNA content for the target and non-target plasmids. Previous studies with other CRISPR–Cas systems that specifically targeted the bacterial chromosome observed similar morphological changes[43,44,45], suggesting that these distinct morphologies are due to dsDNA damage. These results demonstrate that RNA targeting by *Su*Cas12a2 causes dsDNA damage of the bacterial chromosome that in turn induces the SOS response and abortive infection in bacteria, reflecting a distinct mechanism of immunity that relies on indiscriminate dsDNase activity. Consistent with this observation, recent cryo-electron microscopy structures revealed that Cas12a2 binds to and cuts dsDNA through a mechanism that is completely distinct from all other CRISPR-associated nucleases, and structure-guided mutants with impaired in vitro collateral dsDNase but not ssRNase and ssDNase activities abolished in vivo defence activity against plasmids[19].

RNA detection with Cas12a2

CRISPR single-effector nucleases have been repurposed for many applications from gene editing to molecular diagnostics. To determine whether Cas12a2 could be repurposed as a biotechnological tool, we co-opted *Su*Cas12a2 to detect RNA. We programmed apo *Su*Cas12a2 with a crRNA guide complementary to an RNA target and incubated the complex with a ssDNA or ssRNA beacon that fluoresces after cleavage due to separation of a fluorophore and a quencher (Fig. 4f). Using this approach, we were able to detect RNA using both ssDNA and ssRNA probes at 37 °C and room temperature, with a limit of detection within the range observed for other single-subunit Cas nucleases[46] (Fig. 4f and Extended Data Fig. 10a–c). Furthermore, we devised a modified detection assay that uses plasmid DNA and DNA nick translation[47], introducing a distinct positive readout for CRISPR-based diagnostics (Extended Data Fig. 10d–f). These data indicate that Cas12a2 can be readily repurposed as a tool for applications in science, biotechnology, agriculture and medicine. We anticipate that the unique activities of this enzyme can be further leveraged to expand the CRISPR-based toolkit.

Discussion

Collectively, our data support a model in which Cas12a2 nucleases exhibit RNA-triggered degradation of cytoplasmic dsDNA and small RNAs, impeding host cell growth and eliciting an abortive-infection phenotype (Fig. 4h). This mechanism contrasts with targeted invader clearance or abortive-infection activities exhibited by other CRISPR–Cas systems. Specifically, the mechanism exhibited by Cas12a2 is reminiscent of the recently described CBASS defence system, which relies on the indiscriminate double-stranded DNase NucC that degrades host cell DNA and kills the cell[12]. Notably, some type III systems encode NucC enzymes, suggesting that other CRISPR–Cas systems have convergently evolved to use a similar DNA-degrading abortive-infection mechanism[13,48].

In addition to damaging the genome and inducing the SOS response, *Su*Cas12a2 exhibits promiscuous RNA recognition through a flexible PFS

and mismatch tolerance. This flexibility in target recognition mirrors the flexibility of tag–anti-tag complementarity observed with type III and VI systems[8,33] and could be particularly advantageous against rapidly evolving phages, although Cas12a2 orthologues must be characterized to determine whether such promiscuity is a common feature of these nucleases. Although flexible, the PFS would still prevent self-recognition of spurious antisense transcription of the CRISPR array, as the corresponding PFS-containing portion of the repeat strongly diverges from the recognized PFS—a standard feature of self/nonself-recognition for PAMs in DNA-targeting CRISPR–Cas systems[49]. The flexibility in PFS and target recognition could further serve as a back-up mechanism if precise recognition and clearance of DNA targets by Cas12a fails in organisms that encode both Cas12a and Cas12a2 adjacent to a single CRISPR array (Fig. 1a,d and Extended Data Figs. 1 and 6b). This dual-nuclease strategy would be akin to bacteria encoding multiple CRISPR–Cas systems targeting the same invader[50]. However, further research is needed to understand how these two nucleases work together to counter infections.

The combination of nuclease-mediated crRNA biogenesis, RNA targeting and collateral cleavage of ssRNA, ssDNA and, in particular, dsDNA sets Cas12a2 apart from other known Cas nucleases. The apparent need to recognize the A-rich flanking sequence by *Su*Cas12a2 to activate the indiscriminate RuvC nuclease activity strongly indicates that Cas12a2 must bind to a correct PFS adjacent to the RNA target to activate cleavage rather than rely solely on complementarity between the repeat tag and target anti-tag pair to distinguish self from non-self sequences, typical of several other RNA-targeting Cas nucleases and complexes[8,33,51,52]. Investigating the underlying molecular basis of target recognition and activation of collateral cleavage by Cas12a2 could reveal new mechanisms used by CRISPR nucleases to discriminate between self and non-self targets. Recent cryo-electron microscopy structures of Cas12a2 at stages of RNA targeting and collateral dsDNA capture are already fulfilling this need[19].

Cas12a2 holds substantial potential for CRISPR technologies. As a proof-of-principle demonstration, we showed that *Su*Cas12a2 can be repurposed for RNA detection with a limit of detection comparable to existing single-effector based tools[46]. Beyond the ability to detect RNA, we envision a

variety of *Su*Cas12a2 applications that expand and enhance the CRISPR-based tool kit. RNA-triggered dsDNA cleavage could enable programmable killing of prokaryotic and eukaryotic cells with various applications, including programmable shaping of microbial communities, cancer therapeutics and counterselection to enhance genome editing. Moreover, the ability of Cas12a2 and Cas12a to use the same crRNA sequence yet recognize distinct nucleic acid species (RNA versus ssDNA and dsDNA) and elicit distinct non-specific cleavage activities (ssRNA, ssDNA and dsDNA (Cas12a2) versus ssDNA (Cas12a)) could augment existing Cas12a applications by incorporating Cas12a2. By further exploring the properties of *Su*Cas12a2 and its orthologues, we expect the advent of new and improved CRISPR technologies that could broadly benefit society.

Methods

Identification of the putative Cas12a2 nucleases

Several Cas12a2 sequences were initially identified and tentatively classified as encoding Cas12a nucleases[16]. These Cas12a2 protein sequences were used as seeds for BLASTp searches of protein data in NCBI and for tBLASTn searches of metagenomic data in NCBI (https://www.ncbi.nlm.nih.gov) and JGI (https://img.jgi.doe.gov) to identify additional putative Cas12a2 nucleases.

Phylogenetic analysis of Cas12a2 proteins within type V systems

The amino acid sequences of Cas12a2, Cas12a and Cas13b orthologues were aligned using MAFFT (v.7.490)[53]. The resulting alignment was trimmed using ClipKIT[54] and used to create a maximum-likelihood phylogeny using RAxML-NG[55] with the following parameters: --model JTT+G --bs-metric fbp, tbe --tree pars{60}, rand{60} --seed 12345 --bs-trees autoMRE. Cas13b sequences were used as an outgroup. The amino acid sequences used in the creation of the phylogeny are provided in Supplementary File 1.

Domain annotation and structure prediction

Conserved motifs in *Su*Cas12a2 were identified using MOTIF Search (https://www.genome.jp/tools/motif/, accessed on 15 June 2021) and Phyre 2[56] (accessed on 8 March 2021). HHpred secondary structure predictions of Cas12a2 orthologous amino acid sequences were performed to identify the common secondary structure between Cas12a2 and Cas12a that predicted the crRNA processing site of Cas12a2[57].

Strains and plasmids

All of the in vivo experiments, unless indicated otherwise, were performed in *E. coli* BL21(AI). For propagation, the cultures were grown in LB medium at 37 °C with constant shaking at 225–250 rpm. *E. coli* strain TOP10 was used for plasmid cloning (Supplementary Table 1 (tab 1)). All primers, gBlocks and oligos were obtained from Integrated DNA Technologies, unless specified otherwise. Gibson assembly of plasmid construction was performed using the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, E2621). Mutagenesis of the plasmids, including small insertions and nucleotide substitutions, was performed using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs, E0554S). All of the nucleases together with crRNA, unless specified otherwise, were expressed from plasmids containing p15A origin-of-replication and a chloramphenicol-resistance marker. The expression of the nucleases and crRNA was controlled by a T7 promoter, unless otherwise specified. All of the target and non-target plasmids were created by introducing protospacer sequences and corresponding flanking sequences into pBR322 or sc101 origin-of-replication plasmids bearing a kanamycin-resistance cassette, unless otherwise specified. Sequences encoding Cas12a2 orthologues (Supplementary File 1) were codon-optimized and synthesized by Genscript. Sequences encoding *Pb*2Cas12a from *P. bryantii B14* (NCBI: WP_039871282) *Lb*Cas12a from *L. bacterium* ND2006 (NCBI: WP_035635841.1), *Fn*Cas12a from *Francisella tularensis* (NCBI: WP_104928540.1), *As*Cas12a from *Acidaminococcus* sp. BV3L6 (NCBI: WP_021736722.1) and *Mb*3Cas13a from *Moraxella bovoculi* (NCBI: WP_080946945.1)[16] were codon optimized for expression in *E. coli* and ordered as gBlocks from Integrated DNA Technologies. Sequences encoding anti-CRISPR proteins[36,38] (Supplementary Tables 1 and 3) were codon-optimized for expression in *E. coli* and ordered as gBlocks from Integrated

DNA Technologies. The *acr* genes were then PCR-amplified and introduced into the pBAD24 plasmid backbone carrying an ampicillin-resistance cassette[58]. The *Ls*Cas13a-encoding plasmid pCBS2091 was ordered from Addgene (79150)[8]. For detecting RecA-dependent SOS response in *E. coli* BL21(AI), reporter plasmids pCBS2000, pCBS3611 and pCBS3616 were created by introducing the *recA* promoter, included 100 bp upstream of the predicted LexA-binding site, upstream of the GFP-encoding gene into plasmid pCBS198. Plasmids pCBS3611 and pCBS3616 received an ampicillin-resistance cassette from plasmid pCB672[24]. The *recA* promoter sequence was identified in the genome of *E. coli* BL21(AI) between positions 2,635,525 and 2,635,347 (NCBI: CP047231.1). Control plasmids pCBS3616 and pCBS2002 without the GFP-reporter genes were generated by PCR amplification of pCBS2000 and pCBS3616 followed by KLD assembly (New England Biolabs, M0554). A full list of plasmids used in the study, including links to plasmid maps, is provided in Supplementary Tables 1 and 2. A list of relevant oligonucleotide, dsDNA and RNA sequences is provided in Supplementary Tables 1 and 3.

In vitro characterization of *Su*Cas12a2

Expression and purification of *Su*Cas12a2

N-terminal 6×His-tagged *Su*Cas12a2 WT and mutant constructs were expressed in *E. coli* Nico21(DE3) cells from a pACYC plasmid either lacking (apo) (plasmid 1416) or containing a three-spacer CRISPR array (crRNA-guided) (plasmid 1408) using either autoinduction or isopropyl β-d-1-thiogalactopyranoside (IPTG) induction. Autoinduction growth was performed according to previously reported guidelines[59]. In brief, a solution containing recommended concentrations of ZY media, $MgSO_4$, metals mix, 5052 (0.5% glycerol, 0.05% glucose, 0.2% α-lactose) and NPS autoinduction buffers along with antibiotics needed for selection was inoculated with bacteria from a glycerol stock or a fresh transformation. The cells were grown for 5 h at 37 °C with shaking at around 250 rpm and then moved to 24 °C where they were incubated for 24 h before collection by centrifugation at 8,000 rpm for 25 min. Cell pellets were then stored at −80 °C until purification. For the IPTG induction, 1 l of TB medium was

inoculated with 20 ml of overnight growth and was grown at 37 °C until an optical density at 600 nm (OD_{600}) of 0.6. The cells were then cold-shocked on ice for 15 min and induced with 0.1 mM IPTG, followed by a 16–18 h incubation at 18 °C. Cells were collected by centrifugation. Cells were lysed by sonication in lysis buffer (25 mM Tris pH 7.2, 500 mM NaCl, 10 mM imidazole, 2 mM $MgCl_2$, 10% glycerol) in the presence of leupeptin, aprotinin, pepstatin, AEBSF and lysozyme. The lysate was clarified by centrifugation at 36,400*g* for 35 min. Clarified lysate was added to 5 ml of Ni-NTA resin and batch bound at 4 °C for 30 min, and then washed with 100 ml of lysis buffer. The protein was eluted with 50 ml of Ni-elution buffer (25 mM Tris pH 7.2, 500 mM NaCl, 250 mM imidazole, 2 mM $MgCl_2$, 10% glycerol). Fractions containing *Su*Cas12a2 were desalted using the Hiprep 26/10 desalting column into low-salt buffer (25 mM Tris-pCas12a22, 50 mM NaCl, 2 mM $MgCl_2$, 10% glycerol). *Su*Cas12a2 + crRNA was then applied to a Hitrap Q HP column anion-exchange column, whereas the apo *Su*Cas12a2 was applied to a Hitrap SP HP cation-exchange column. The column was washed with 10% high-salt buffer (25 mM Tris pH 7.2, 1 M NaCl, 2 mM $MgCl_2$, 10% glycerol) followed by a gradient elution to 100% high-salt buffer 10 CV (50 ml). The fractions containing *Su*Cas12a2 were concentrated using a 100 kDa MWKO concentrator to about 1 ml and then purified by size-exclusion column chromatography over a Hiload 26/600 Superdex 200 pg column equilibrated in SEC buffer (100 mM HEPES pH 7.2, 150 mM KCl, 2 mM $MgCl_2$, 10% glycerol). The fractions containing *Su*Cas12a2 were concentrated and stored at −80 °C.

Pre-crRNA processing

For processing of a 3× pre-crRNA, *Su*Cas12a2 pre-CRISPR×3 RNA was in vitro transcribed using the HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs). The template DNA was derived from Jackson Laboratory plasmid 1409 linearizing with the KpnI restriction enzyme. A contaminating band that runs approximately at 130 nucleotides was observed to be an artifact of the reaction. Numerous strategies were attempted to prevent the transcription of this contaminating band, to no success. In vitro transcribed RNA was cleaned using RNeasy spin columns (Qiagen). Then, 1.5 µM of apo *Su*Cas12a2 was incubated with 1 mg of *Su*Cas12a2 pre-CRISPR×3

RNA in 1× 3.1 buffer from New England Biolabs (100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 100 mg ml^{-1} BSA pH 7.9) and incubated at 25 °C for various times. The samples were run on a gel (12% polyacrylamide, 8 M, TBE) alongside a ssRNA low range ladder (New England Biolabs) and stained with SYBR gold (Thermo Fisher Scientific).

For processing of a 1× crRNA with WT and crRNA-processing mutants, a synthetic crRNA with a 13 base 5′ unprocessed overhang (smcrRNA; Supplementary Tables 1 and 3) was refolded using a previously outlined protocol[60]. In a 10 µl reaction, 150 nM of crRNA substrate was combined with 1.5 µM WT, K784A or K785A apo *Su*Cas12a2 protein in NEB 3.1. The reactions were incubated at 37 °C for 1 h. The reactions were quenched with phenol and phenol–chloroform extraction was performed. The results were analysed using 12% urea–PAGE stained with SYBR Gold.

Nucleic acid cleavage assays

For analysis of targeted cleavage, 10 µl reactions of 250 nM *Su*Cas12a2–crRNA with 100 nM of complementary FAM-labelled synthetic oligonucleotide (that is, ssDNA, dsDNA or RNA) in 1× NEB 3.1 buffer were incubated at 37 °C for 1 h. Reactions were quenched with phenol and then phenol–chloroform extraction was performed. Results were analysed using a previously outlined FDF-PAGE method[61] and visualized for fluorescein fluorescence.

For analysis of collateral cleavage, 10 µl reactions of 250 nM *Su*Cas12a2–crRNA, and 250 nM of target (RNA complementary to the crRNA-guide) or non-target (RNA non-complementary to the crRNA-guide) substrate and 100 nM of 5′-FAM labelled collateral substrate (ssDNA, dsDNA, RNA) in 1× NEB 3.1 were incubated at 37 °C for 1 h. Reactions were quenched with phenol and then phenol–chloroform extraction was performed. The results were analysed using 12% urea–PAGE and visualized for fluorescein fluorescence.

For analysis of the flanking-sequence requirements for activation, 10 µl reactions of 250 nM Cas12a2–crRNA, with 300 nM of different target ssRNAs (self (flanked by sequence complementary to the direct repeat of the

crRNA), no flanks and flanks containing a 5′-GAAA-3′ PFS on the 3′ side of the protospacer) and 100 nM of collateral 5′-FAM dsDNA in 1× NEB 3.1 buffer were incubated at 37 °C for 1 h. The reactions were quenched with phenol and phenol–chloroform extraction was performed. The results were analysed using 12% urea–PAGE and visualized for fluorescein fluorescence.

For kinetic analysis of collateral cleavage, a single 100 µl reaction containing 100 nM Cas12a2–crRNA, 100 nM of target ssRNA (crRNA complementary) and 100 nM of different 5′-FAM labelled collateral substrates (ssDNA, dsDNA, RNA) in 1× NEB 3.1 buffer was made. Time points were taken at 1, 2, 5, 10, 15, 30, 60, 120 and 180 min by combining 10 µl from the 100 µl reaction with phenol, and then performing phenol–chloroform extraction. The results were analysed using 12% urea–PAGE and visualized for fluorescein fluorescence.

For the plasmid-cleavage assay, a 100 µl reaction containing 14 nM Cas12a2–crRNA, 25 nM target RNA, 7 nM of pUC19 plasmid in 1× NEB 3.1 buffer was incubated at 37 °C. At the indicated time points, 10 µl of the reaction was removed and quenched with phenol and phenol–chloroform extraction was performed. The reactions were visualized on 1% agarose with ethidium bromide.

Collateral cleavage comparison between Cas12a2 and Cas12a, Cas13a and Cas12g

EnGen LbaCas12a (*Lb*Cas12a) was purchased from New England Biolabs (M0653S). Reactions (10 µl) containing 250 nM of *Lb*Cas12a and 500 nM of its cognate crRNA in 1× NEB 2.1 buffer were incubated at 37 °C with 200 nM of different target substrates (ssDNA, dsDNA, RNA) and 100 nM of different FAM-labelled collateral substrates (ssDNA, dsDNA, RNA). After 1 h, the reactions were quenched by phenol and phenol–chloroform extraction was performed. The results were analysed using 12% urea–PAGE and visualized for fluorescein fluorescence.

*Lw*Cas13a was purchased from MCLAB Molecular Cloning Laboratories (Cas13a-100). Reactions (10 µl) containing 250 nM of *Lw*Cas13a and 500 nM of its cognate crRNA in the provided 1× Cas9 buffer (20 mM HEPES

(pH 6.5), 5 mM $MgCl_2$, 100 mM NaCl, 100 µM EDTA) were incubated at 37 °C with 200 nM of different target substrates (ssDNA, dsDNA, RNA) and 100 nM of different FAM-labelled collateral substrates (ssDNA, dsDNA, RNA). After 1 h, the reactions were quenched by phenol and phenol–chloroform extraction was performed. The results were analysed using 12% urea–PAGE and visualized for fluorescein fluorescence.

*Ab*Cas12g was expressed in *E. coli* NiCo 21 DE3 using pET28a-mH6-Cas12g1 (Addgene plasmid, 120879) and initially purified as described previously[25]. The protein was then transferred to low-salt buffer (25 mM HEPES pH 7.8, 50 mM NH_4Cl, 2 mM $MgCl_2$, 7 mM BME, 5% glycerol) by buffer-exchange and loaded over heparin followed by elution with a linear NaCl gradient and gel-filtration as described previously[62]. Purified protein was flash-frozen and stored at −80 °C. The Cas12g1 non-coding plasmid pACYC-Cas12g1 (Addgene plasmid, 120880) was used as a template for PCR amplification of the *Ab*Cas12g tracrRNA sequence with Cas12gtracrRNA F and R primers (Supplementary Tables 1 and 3) in 2× Taq Master Mix (New England Biolabs). The non-coding plasmid was removed with DpnI by incubation at 37 °C for 1 h in CutSmart buffer (New England Biolabs). DNA components were cleaned after PCR and DpnI digestion using the E.Z.N.A. Cycle Pure Kit (OMEGA BioTek). The Cas12g tracrRNA was transcribed using the HighScribe T7 Quick High Yield RNA synthesis kit and cleaned using the Monarch RNA cleanup kit (New England Biolabs). Reactions (10 µl) containing 250 nM of Cas12g, 500 nM of the Cas12g crRNA and 1 µM of Cas12g tracrRNA in 1× NEB 3.1 buffer were incubated at 37 °C or 50 °C with 200 nM of different target substrates (ssDNA, dsDNA, RNA) and 100 nM of different FAM-labelled collateral substrates (ssDNA, dsDNA, RNA). After 1 h, the reactions were quenched by phenol and phenol–chloroform extraction was performed. The results were analysed using 12% urea–PAGE and visualized for fluorescein fluorescence.

To analyse *Su*Cas12a2 collateral activity, 10 µl reactions containing 250 nM of Cas12a2–crRNA, 200 nM of different target substrates (ssDNA, dsDNA, ssRNA) and 100 nM of different FAM-labelled collateral substrates (ssDNA, dsDNA, RNA) in 1× NEB 3.1 buffer were incubated at 37 °C for 1 h. The reactions were quenched with phenol and phenol–chloroform extraction was

performed. The results were analysed using 12% urea–PAGE and visualized for fluorescein.

RNA detection by Cas12a2 with ssRNA and ssDNA reporter probes

Cas12a2 (100 nM) was complexed with crRNA (120 nM) in NEB 3.1 buffer (50 mM Tris-HCl pH 7.9, 100 mM NaCl, 10 mM $MgCl_2$, 100 μg ml^{-1} BSA) before combining with RNase or DNase Alert (200 nM, IDT) and Target RNA to the indicated concentrations in a 384-well plate (Greiner Bio-One, 784077). A background control was prepared with nuclease-free water instead of target RNA. The reactions were monitored for reporter fluorescence (RNase Alert: excitation 485-20/emission 528-20, DNAse Alert: excitation 500-20/emission 560-20) over time at either ambient conditions (room temperature) or 37 °C using the Synergy H4 Hybrid multi-mode microplate reader (BioTek Instruments). The slope of the linear region (between 5 and 30 min) was determined at each concentration of target RNA using GraphPad PRISM. Standard error of the linear fit was used as a proxy for standard deviation, and the limit of detection was calculated as 3 × standard error of the water background as described previously[46]. The limit of detection was estimated by determining where the plot of V_0 (relative fluorescence units (RFU)/s) versus concentration of target RNA crosses the detection threshold.

Plasmid cleavage

Plasmid cleavage reactions were prepared by combining 14 nM *Su*Cas12a2 (or mutant) with 14 nM crRNA and 25 mM Target RNA in NEB 3.1 buffer (50 mM Tris-HCl pH 7.9, 100 mM NaCl, 10 mM $MgCl_2$, 100 μg ml^{-1} BSA). Protein was preheated at 37 °C for 15 min before the addition of 7 nM supercoiled pUC19 plasmid. Samples were removed at the time points 1, 2, 5, 10, 20, 30, 45 and 60 min and quenched in pH 8.0 phenol–chloroform. Quenched reactions were mixed by flicking followed by centrifugation. The samples were loaded onto 1% agarose gels and visualized using ethidium bromide. Gels were imaged using the ChemiDoc MP Image System (Bio-Rad).

Nick translation

The plasmid pSPC421 was collected from TOP10 *E. coli* cells using the ZymoPURE II Plasmid Midiprep Kit by Zymo Research (D4201), and cleaned using the DNA Clean & Concentrator-5 kit from Zymo Research (D4013). ca33Cas12a2 was expressed from the plasmid pCBS5042 and purified as described above at the Rudolf Virchow Center for Integrative and Translational Bioimaging. ca33Cas12a2 nuclease (100 nM) was incubated with the crRNA (1 µM) in the NEB3.1 buffer for 30 min at room temperature. CAO1 target RNA (1 nM) and pSPC425 (3 µg) were added to the reaction medium for 15 min. For evaluating plasmid nicking, the samples were heated at 80 °C for between 1 and 30 min. The reactions were run on a 0.8% agarose gel. DNA polymerase I (NEB, M0209L) was added to the reaction (0.2 U μl^{-1}) with the Atto421-NT labelling mix (1×) and NT labelling buffer (1×) from the Atto425 NT Labelling Kit (Jena Bioscience, PP-305S-425). The samples were incubated at 15 °C for 90 min in a Bio-Rad thermocycler. The resulting labelled DNA fragments were purified using the Microspin S-400 HR columns (Cytiva, 27514001). Fluorescence measurements (λ_{exc} = 436 nm; λ_{em} = 484 nm) were performed on a fluorescence microtitre plate reader (BioTek NeoG2) at 25 °C.

Cas12a2 characterization in *E. coli*

crRNA sequencing and analysis

The *Su*Cas12a2 expression plasmid pCBS3568 containing the nuclease- and the crRNA-encoding sequences and the no-crRNA control pCBS3569 were transformed into *E. coli* BL21(AI) and the transformants were plated on selection pates. The resulting colonies were picked and used to inoculate 2 ml overnight liquid cultures. The next day, the overnight cultures were used to inoculate 25 ml of LB containing chloramphenicol to an OD_{600} of approximately 0.05. Once the growing cultures reached an OD_{600} of 0.25 after approximately 40 min, expression of the nuclease and the crRNA were induced with 1 mM IPTG and 0.2% l-arabinose. The induced cultures were collected in the stationary phase by centrifugation at 14,000 rpm at 4 °C for

2 min. The cell pellets were then immediately frozen in liquid N_2 and stored at −80 °C until further processing.

Total RNA was purified from cell pellets using the Direct-zol RNA Miniprep Plus (Zymo Research, R2072) according to the manufacturer's instructions. DNA was removed using Turbo DNase (Life Technologies, AM2238). Between the individual processing steps, RNA was purified using the RNA Clean & Concentrator kit (Zymo Research, R1017). Ribosomal RNA was removed from the samples using the RiboMinus Transcriptome Isolation Kit, bacteria (Thermo Fisher Scientific, K155004). 3′-phosphoryl groups were removed from RNA using T4 polynucleotide kinase (New England Biolabs, M0201S). cDNA synthesis and library preparation was performed using the NEBNext Multiplex Small RNA Library Prep Set for Illumina (New England Biolabs, E7330S). Size selection for fragments between 200 bp and 700 bp was performed using the Select-a-Size DNA Clean & Concentrator kit (Zymo Research, D4080). Finally, DNA was purified using AMPure XP beads (Beckman Coulter, A63882) and quantified using the Qubit dsDNA HS assay kit (Thermo Fisher Scientific, Q32851) on DeNovix DS-11 FX (DeNovix).

Library sequencing was performed at the Helmholtz Center for Infectious Research (HZI) GMAK facility in Braunschweig, Germany, using the MiSeq 300 sequencing method (Illumina). The resulting paired-end reads were quality controlled, trimmed and merged using BBTools[63] (https://sourceforge.net/projects/bbmap/). Next, the reads were mapped to the crRNA expression site on the plus strand of pCBS273 using Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/). The associated raw and processed sequencing data as well the data-processing steps can be found at the NCBI Gene Expression Omnibus (GEO: GSE178531).

Plasmid clearance assay in *E. coli*

Standard plasmid clearance assays were performed in *E. coli* BL21(AI) containing nuclease- and crRNA-expressing plasmids. Bacterial cultures were grown overnight and used to inoculate fresh LB medium containing chloramphenicol to an OD_{600} of 0.05–0.1. Subsequently, these cultures were grown until the OD_{600} reached approximately 0.25, at which time 1 mM

IPTG and 0.2% l-arabinose were added for induction. Once the cultures reached an OD_{600} of 0.6–0.8, the cells were collected and made electrocompetent[64]. Electrocompetent cells were prepared from four biological replicates. Immediately after, 1 µl of 50 ng μl^{-1} of the target and non-target plasmid were electroporated into 50 µl of the electrocompetent *E. coli* cells. To achieve high transformation efficiencies, the used plasmids were purified through ethanol precipitation and quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Q32851). The electroporated cells were recovered for 1 h at 37 °C with shaking in 500 µl LB containing 1 mM IPTG and 0.2% l-arabinose without antibiotics. Next, the cultures were sequentially diluted to 10^{-5} in tenfold increments. Then, 5–10 µl of each dilution was spotted onto LB plates containing antibiotics to select the nuclease-crRNA and the target/non-target plasmids. The plates also contained 0.3 mM IPTG and 0.2% l-arabinose. The plates were incubated overnight at 37 °C.

The next day, the colonies were manually counted and the resulting counts were adjusted for the dilution factor. Counts from the highest countable dilution were used to calculate transformation fold reduction as a ratio between the colonies in the non-target condition divided by the colonies in the target condition.

In a modification of the assay used to determine the cell suicide phenotype, the target and the non-target plasmids were transformed into *E. coli* BL21(AI) first. Next, these cells were made electrocompetent and the nuclease-crRNA plasmids were transformed in last.

When testing Acrs, the Acr plasmid (ampicillin) and the nuclease-crRNA plasmid (chloramphenicol) were co-transformed, followed by electroporation of the target or non-target plasmid (kanamycin).

Growth experiments

To investigate the growth of the cultures under nuclease-targeting conditions, the nuclease-crRNA and the target/non-target plasmids were transformed into *E. coli* BL21(AI). The resulting transformants were recovered in SOC medium and grown overnight with 0.2% glucose to inhibit

nuclease and crRNA expression. In the morning, the cells were collected by centrifugation at 5,000*g* for 2 min. The pellets were resuspended in LB and used to inoculate 200 µl of LB medium on a 96-well plate to a final OD_{600} of 0.01. Depending on the experiment, the reactions contained different combinations of antibiotics, IPTG and l-arabinose. The plates were incubated in the BioTek Synergy H1 plate reader at 37 °C with vigorous shaking. The OD_{600} of the cultures was recorded every 3 min. Plasmid clearance assays were performed with the overnight cultures, as described above.

PFS depletion assay in *E. coli*

To determine PFS preferences of *Su*Cas12a2, a PFS depletion assay was performed. An oligo library (ODpr23) consisting of 1,024 nucleotide combinations in place of a 5-nucleotide PFS-encoding site was synthesized by Integrated DNA Technologies. Using the ODpr23 oligo pool library in a combination with primer ODpr24, targeting plasmid pCBS276 was PCR-amplified using Q5 polymerase (New England Biolabs, M0543). The PCR products were gel-purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research, D4007) and ligated using the KLD reaction mix (New England Biolabs, M0554). The ligated plasmids were purified using ethanol precipitation and electroporated into *E. coli* TOP10. A total of ten electroporation reactions were performed. After recovery of the electroporated cells in SOC medium, the individual reactions were combined to inoculate 90 ml of LB medium containing kanamycin. A total of 10 µl from each electroporation reaction was plated on selective LB medium to estimate the total number of transformed bacteria. With the colony counts, we estimated that the total number of transformed cells exceeded the number of unique PAM sequences in the library (1,024) by approximately 2,300-fold. Plasmid library DNA was purified from the combined overnight culture using the ZymoPURE II Plasmid Midiprep Kit (Zymo Research, D4201) and additionally cleaned by ethanol precipitation. Next, the plasmid library was verified by Sanger sequencing.

The PAM plasmid library was transformed into electrocompetent *E. coli* BL21(AI) containing either the *Su*Cas12a2 nuclease-expressing plasmid pCBS273 or an empty plasmid control pCBS3569. The electrocompetent

cells were prepared as described above. Approximately 600 ng of the plasmid DNA was electroporated into 50 µl volume of the competent cells. The transformed bacteria were recovered in 500 µl of SOC medium for 1 h at 37 °C and were used to inoculate 50 ml LB with 1 mM IPTG and 0.2% l-arabinose in the presence of kanamycin and chloramphenicol. The cultures were grown for 13 h before the cells were collected by centrifugation at 4,000*g* for 15 min and the plasmid DNA extracted using ZymoPURE II Plasmid Midiprep Kit (Zymo Research, D4201). After recovery, bacteria were also plated on LB plates containing kanamycin and chloramphenicol without the inducers. These plates were used to estimate the total number of cells transformed with the plasmid library. The total number of transformed cells estimated based on the colony counts exceeded the number of unique PAM sequences in the library by approximately 1,700-fold for the cells containing the *Su*Cas12a2–crRNA plasmid (pCBS273) compared to 11,900-fold in the no-crRNA control (pCBS3569).

The region of the plasmid DNA containing the target site including the PFS-encoding sequence was PCR-amplified using the primers ODpr55 and ODpr56. The PCR reactions were purified using AMPure XP beads (Beckman Coulter, A63882). The purified PCR products were indexed using the primers ODpr58, ODpr60, ODpr59 and ODpr61. The indexed PCR products were purified using the AMPure XP beads, quantified using the Qubit assay (Thermo Fisher Scientific, Q32851) and sent for sequencing at the HZI GMAK facility using the MiSeq PE300 Illumina sequencing method.

Analysis of the PFS-encoding sequence depletion data as well as the creation of the PFS wheels were performed as described previously[65]. PFS consensus motifs were defined manually. The raw and the processed sequencing data as well as the data-processing steps can be found at the NCBI GEO (GSE178530). Individual PFS sequences were validated using plasmid-clearance assays as described above.

Cell-free transcription–translation reactions

For in vitro assays to test Acr sensitivity of Cas12a nucleases, plasmids encoding Cas12a nuclease were pre-expressed together with a plasmid

encoding either a target or non-target crRNA in 9 µl of MyTXTL master mix (Arbor Biosciences) at a final concentration of 4 nM for each plasmid in a total volume of 12 µl. Acrs were pre-expressed separately, at a concentration of 4 nM in a total volume of 12 µl. As the Acrs are encoded on linear DNA fragments, GamS at a final concentration of 2 µM was added to prevent DNA degradation. All pre-expressions were performed at 29 °C for 16 h. The subsequent cleavage assay was performed by adding 1 µl of each pre-expression reaction to 9 µl of fresh myTXTL mix. pCBS420 plasmid constitutively expressing deGFP protein was used as a reporter at a final concentration of 1 nM. For quantification, four 3 µl replicates per reaction were transferred onto a 96-well V-bottom plate (Corning Costar 3357). The reactions were prepared using the Echo 525 Liquid Handler (Beckman Coulter). Fluorescence was measured on the BioTek Synergy H1 plate reader (excitation, 485/20; emission, 528/20). Time-course measurements were run for 16 h at 29 °C, with 3 min intervals between the measurements.

All fold-repression values for plasmid reporter constructs represent the ratio of deGFP concentrations after 16 h of reaction for the non-target over the target crRNA. For the experiments measuring the inhibitory activity of Acrs, inhibition was calculated from end-point expression values after 16 h of expression according to the following formula[66]: percentage inhibition of nuclease activity = $100 \times (RFU_{t,Acr}/RFU_{nt,Acr} - RFU_{t,-}/RFU_{nt,-})/(1 - RFU_{t,-}/RFU_{nt,-})$, where the inhibition of nuclease activity (%) is defined by the ratio of fluorescence between GFP targeting (t) and non-targeting (nt) Cas nucleases in the presence (Acr) and absence (-) of Acrs.

Quantification of SOS response

To measure the RecA-dependent SOS response, the nuclease-crRNA, the target/non-target plasmids (pCBS276/pCBS3578, kanamycin) and the reporter PrecA-gfp/no-gfp (pCBS3611/pCBS3616, ampicillin) plasmids were transformed into *E. coli* BL21(AI) sequentially. The plasmids pCBS273 and pCBS3588 (chloramphenicol) were used to express *Su*Cas12a2 and *Lb*Cas12a nucleases, respectively. When measuring the RecA-dependent SOS response in the presence of *Ls*Cas13a, the nuclease expression plasmid pCBS361 (chloramphenicol) was used. The target/non-target plasmids pCBS2004/pCBS612 (ampicillin) and PrecA-gfp/no-gfp

plasmids pCBS2000/pCBS2002 (kanamycin) were used. First, the cells were grown in LB medium with 0.2% glucose to inhibit the expression of the nucleases and the crRNA. The bacteria were collected from the overnight cultures (15 ml) by centrifugation at 5,000*g* for 2 min and resuspended in fresh LB. Next, 200 µl of fresh LB medium was inoculated onto 96-well plates with the resuspended bacteria from the overnight cultures. These cultures were grown in the presence of either chloramphenicol, kanamycin and ampicillin, chloramphenicol and ampicillin, or no antibiotics. For the induction of nuclease and crRNA expression 1 mM of IPTG and 0.2% l-arabinose were added.

The cultures were grown at 37 °C with vigorous shaking. OD_{600} and fluorescence measurements (excitation: 485/20, emission: 528/20) were collected every 5 min on the BioTek Synergy H1 plate reader. Four biological replicates were measured per experimental condition.

To determine whether a change in fluorescence occurred as a result of nuclease targeting, first the background fluorescence collected for the cultures with the PrecA-no-gfp plasmid (pCBS3616/pCBS2002) was subtracted from the values obtained for the cultures with the GFP-expressing plasmids for each time point (pCBS3611/pCBS2000). Next, the fluorescence values were divided by the OD_{600} values from the corresponding target and the non-target cultures. Statistical significance was determined using Welch's *t*-test with unequal variance.

In parallel we performed a plasmid-clearance assay with the washed overnight cultures (Supplementary Fig. 9b), as described above. For the lowest plated dilution, cultures at an OD_{600} of around 0.1 were used.

Flow cytometry

For the flow cytometry measurements, *E. coli* BL21(AI) cells were sequentially electroporated with the nuclease-encoding and target/non-target plasmids. The *Su*Cas12a2- and *Lb*Cas12a-expressing plasmids pCBS273 and pCBS3588 were used, respectively. Target plasmid pCBS273 and non-target plasmid pCBS3578 were used. For the experiments involving *Ls*Cas13a, nuclease-expression plasmid pCBS361 was used in combination with the

target plasmid pCBS2004 and non-target plasmid pCBS612. After plasmid transformation, the *E. coli* bacteria were recovered in SOC medium and grown overnight in LB with chloramphenicol, kanamycin and 0.2% glucose. Next, the cells were collected at 5,000*g* for 2 min and resuspended in fresh LB. The resuspended bacteria were used to inoculate 15 ml cultures to an OD_{600} of about 0.01. These cultures were grown at 37 °C with 220 rpm shaking for 6 h without antibiotics with 1 mM IPTG and 0.2% l-arabinose. Every 2 h the OD_{600} of the cultures was measured and 500 µl samples were collected and centrifuged for 3 min at 5,000*g*. The cell pellets were then resuspended in 1× PBS containing 2 µg ml^{-1} DAPI (Thermo Fisher Scientific, 62248). The resuspended cells were stained for 10 min in the dark, after which 10 µl was transferred into 240 µl of 1× PBS on a 96-well plate. DAPI fluorescence was measured using the Cytoflow Novocyte Quanteon flow cytometer as emission in the Pacific Blue spectrum (455 nm). Data regarding the forward scatter (FSC) and the side scatter were also collected.

The resulting data were analysed in Python. First, clusters of bacteria exhibiting distinct FSC and Pacific Blue signals were identified using density-based spatial clustering of applications with noise (DBSCAN; https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html). Next, the ratios of the Pacific Blue to the FSC signal for each data point and the percentage of the data points within each cluster were parsed from the clustering data. The resulting values were plotted in the form of balloon plots. A total of 60,000 events per sample were analysed.

Dead/live staining

Dead and viable bacteria were estimated using the LIVE/DEAD BacLight Bacterial Viability and Counting Kit (Molecular Probes, L34856). The measurements were performed using the Cytoflow Novocyte Quanteon flow cytometer. *E. coli* BL21(AI) bacteria were transformed with nuclease, crRNA, and either target or non-target expression plasmids. For expressing *Su*Cas12a2 and *Lb*Cas12a with a target guide, the plasmids pCBS273 and pCBS3588 were used, respectively. Target expression plasmid pCBS2004 and non-target expression plasmid pCBS612 were used. For expressing

*Ls*Cas13a with target and non-target guides, plasmids pCBS273 and pCBS3578 were used, respectively. Cultures containing combinations of nuclease–guide and target plasmids were grown for approximately 16 h with 0.2% glucose inhibitor in four biological replicates. Next, 1 ml of each culture was collected by centrifugation at 5,000*g* for 3 min. The resulting pellet was resuspended in 1 ml of fresh LB medium. A total of 60 µl of this suspension was used to inoculate 20 ml of LB. Three cultures were grown for 2 h at 37 °C with constant shaking at 220 rpm. The expression of the nucleases and the guides was induced with 0.2% arabinose and 0.01 mM IPTG. After 4 h, the OD_{600} of the cultures was measured. A volume of the cultures corresponding to an OD_{600} of 1.0 was collected and processed as described in the kit manual. In brief, samples of the bacterial culture were centrifuged at 10,000*g* for 3 min to pellet the cells. The supernatant was removed and the pellet was resuspended in 1 ml of 0.85% NaCl. As a control for the dead cells, spectate pellet was first resuspended in 300 µl 0.85% NaCl and then 700 µl 70% isopropyl alcohol (dead-cell suspension). The samples were incubated at room temperature for 60 min, with mixing every 15 min. Next, the samples were centrifuged at 10,000*g* for 3 min and washed in 1 ml 0.85% NaCl, followed by another centrifugation. Finally, the samples were resuspended in 0.5 ml of 0.85% NaCl. One millilitre of the master mix for staining the cells contained 977 µl of 0.85% NaCl, 1.5 µl of component A (3.34 mM SYTO 9 nucleic acid stain), 1.5 µl of component B (30 mM propidium iodide (PI)), 10 µl of component C (beads) and 10 µl of the sample. These reactions were incubated for 15 min at room temperature protected from light. Fluorescence was collected in the green (fluorescein for SYTO 9) and red (Texas Red for PI) channels. The dead cells in each sample were gated on the basis of the dead-cell suspension control treated with isopropyl alcohol. The percentage of dead cells stained with PI was calculated from the total number of events without the beads. A total of 50,000 events were counted per sample.

In vivo RNA degradation

Samples corresponding to 1 ml of culture at an OD_{600} of 0.4 grown for dead/live staining, as described above, were collected and centrifuged at 10,000*g* for 3 min. The resulting pellets were frozen in liquid nitrogen and stored −80 °C until further processing. Total RNA was extracted using 1.5

ml of Trizol and 1.5 ml of ethanol with the Direct-zol RNA Miniprep kit (R2051, Zymo), according to the manufacturer's instructions. The RNA was further purified using the RNA Clean & Concentrator-5 kit (R1013, Zymo). A total of 0.5 µg of RNA from each sample in 5 µl was combined with 2.5 µl of RNA loading dye, heated to 70 °C for 10 min and subsequently chilled on ice for 2 min. The RNA High-Range ladder that was used was also heat-treated. The denatured samples (5 µl) and the leader (3 µl) were loaded onto a 1% TBE gel. The gel was run for 40 min at 120 V. Next, the gel was stained for 30 min in ethidium bromide, washed for 10 ml and imaged. Gel images were analysed using GelAnalyzer v.19.1 (www.gelanalyzer.com).

Microscopy

For confocal microscopy, the cells were grown as described above for flow cytometry. At 2 h intervals, 500 µl of each culture was collected and centrifuged at 5,000*g* for 3 min. Next, the bacteria were diluted to approximately the same cell density and stained with 2 $\mu g\ ml^{-1}$ of FM4-64 dye (Thermo Fisher Scientific, T13320) and 1 $\mu g\ ml^{-1}$ of DAPI (Thermo Fisher Scientific, 62248). Imagining was performed on the Leica DMi6000B TCS-SP5 II Inverted Confocal Microscope at ×1,000 magnification.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The NGS data from the PAM depletion assay and crRNA sequencing data were deposited at the NCBI GEO under accession code GSE178536. All other data supporting the findings in the Article and the Supplementary Information are available from the corresponding authors on reasonable request.

References

1. Lopatina, A., Tal, N. & Sorek, R. Abortive infection: bacterial suicide as an antiviral immune strategy. *Annu. Rev. Virol.* **7**, 371–384 (2020).

2. Koonin, E. V. & Krupovic, M. Origin of programmed cell death from antiviral defense? *Proc. Natl Acad. Sci. USA* **116**, 16167–16169 (2019).

3. Meeske, A. J., Nakandakari-Higa, S. & Marraffini, L. A. Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241–245 (2019).

4. Rostøl, J. T. & Marraffini, L. A. Non-specific degradation of transcripts promotes plasmid clearance during type III-A CRISPR–Cas immunity. *Nat. Microbiol.* **4**, 656–662 (2019).

5. Rostøl, J. T. et al. The Card1 nuclease provides defence during type III CRISPR immunity. *Nature* **590**, 624–629 (2021).

6. VanderWal, A. R., Park, J.-U., Polevoda, B., Kellogg, E. H. & O'Connell, M. R. CRISPR-Csx28 forms a Cas13b-activated membrane pore required for robust CRISPR-Cas adaptive immunity. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.02.466367 (2021).

7. Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438–442 (2017).

8. Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).

9. Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605–609 (2017).

10. Niewoehner, O. et al. Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543–548 (2017).

11. Makarova, K. S., Anantharaman, V., Grishin, N. V., Koonin, E. V. & Aravind, L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front. Genet.* **5**, 102 (2014).

12. Lau, R. K. et al. Structure and mechanism of a cyclic trinucleotide-activated bacterial endonuclease mediating bacteriophage immunity. *Mol. Cell* **77**, 723–733 (2020).

13. Grüschow, S., Adamson, C. S. & White, M. F. Specificity and sensitivity of an RNA targeting type III CRISPR complex coupled with a NucC endonuclease effector. *Nucleic Acids Res.* **49**, 13122–13134 (2021).

14. Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).

15. Marino, N. D., Pinilla-Redondo, R. & Bondy-Denomy, J. CRISPR-Cas12a targeting of ssDNA plays no detectable role in immunity. *Nucleic Acids Res.* **50**, 6414–6422 (2022).

16. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).

17. Jinek, M. et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).

18. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Classification and nomenclature of CRISPR-Cas systems: where from here? *CRISPR J.* **1**, 325–336 (2018).

19. Bravo, J. P. K. et al. RNA targeting unleashes indiscriminate nuclease activity of CRISPR–Cas12a2. *Nature* https://doi.org/10.1038/s41586-022-05560-w (2022).

20. Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–521 (2016).

21. Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).

22. Swarts, D. C., van der Oost, J. & Jinek, M. Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR-Cas12a. *Mol. Cell* **66**, 221–233 (2017).

23. East-Seletsky, A. et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**, 270–273 (2016).

24. Marshall, R. et al. Rapid and scalable characterization of CRISPR technologies using an *E. coli* cell-free transcription-translation system. *Mol. Cell* **69**, 146–157 (2018).

25. Yan, W. X. et al. Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88–91 (2019).

26. Murugan, K., Seetharam, A. S., Severin, A. J. & Sashital, D. G. CRISPR-Cas12a has widespread off-target and dsDNA-nicking effects. *J. Biol. Chem.* **295**, 5538–5553 (2020).

27. Collias, D. & Beisel, C. L. CRISPR technologies and the search for the PAM-free nuclease. *Nat. Commun.* **12**, 555 (2021).

28. Elmore, J. R. et al. Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev.* **30**, 447–459 (2016).

29. Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, Č. & Siksnys, V. Spatiotemporal control of type III-A CRISPR-Cas immunity: coupling DNA degradation with the target RNA recognition. *Mol. Cell* **62**, 295–306 (2016).

30. Liu, T. Y., Iavarone, A. T. & Doudna, J. A. RNA and DNA targeting by a reconstituted *Thermus thermophilus* type III-A CRISPR-Cas system. *PLoS ONE* **12**, e0170552 (2017).

31. Han, W. et al. A type III-B CRISPR-Cas effector complex mediating massive target DNA destruction. *Nucleic Acids Res.* **45**, 1983–1993 (2017).

32. Wang, B. et al. Structural basis for self-cleavage prevention by tag:anti-tag pairing complementarity in type VI Cas13 CRISPR systems. *Mol. Cell* **81**, 1100–1115 (2021).

33. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568–571 (2010).

34. Wiedenheft, B. et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl Acad. Sci. USA* **108**, 10092–10097 (2011).

35. Semenova, E. et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl Acad. Sci. USA* **108**, 10098–10103 (2011).

36. Watters, K. E., Fellmann, C., Bai, H. B., Ren, S. M. & Doudna, J. A. Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science* **362**, 236–239 (2018).

37. Knott, G. J. et al. Broad-spectrum enzymatic inhibition of CRISPR-Cas12a. *Nat. Struct. Mol. Biol.* **26**, 315–321 (2019).

38. Marino, N. D. et al. Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science* **362**, 240–242 (2018).

39. Dong, L. et al. An anti-CRISPR protein disables type V Cas12a by acetylation. *Nat. Struct. Mol. Biol.* **26**, 308–314 (2019).

40. Little, J. W. & Mount, D. W. The SOS regulatory system of *Escherichia coli*. *Cell* **29**, 11–22 (1982).

41. Janion, C. Some aspects of the SOS response system-a critical survey. *Acta Biochim. Pol.* **48**, 599–610 (2001).

42. Chen, Z., Lu, M., Zou, D. & Wang, H. An *E. coli* SOS-EGFP biosensor for fast and sensitive detection of DNA damaging agents. *J. Environ. Sci.* **24**, 541–549 (2012).

43. Vercoe, R. B. et al. Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet.* **9**, e1003454 (2013).

44. Cui, L. & Bikard, D. Consequences of Cas9 cleavage in the chromosome of *Escherichia coli*. *Nucleic Acids Res.* **44**, 4243–4251 (2016).

45. Caliando, B. J. & Voigt, C. A. Targeted DNA degradation using a CRISPR device stably carried in the host genome. *Nat. Commun.* **6**, 6989 (2015).

46. Huyke, D. A. et al. Enzyme kinetics and detector sensitivity determine limits of detection of amplification-free CRISPR-Cas12 and CRISPR-Cas13 diagnostics. *Anal. Chem.* **94**, 9826–9834 (2022).

47. Rigby, P. W., Dieckmann, M., Rhodes, C. & Berg, P. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *J. Mol. Biol.* **113**, 237–251 (1977).

48. Nemudraia, A. et al. Sequence-specific capture and concentration of viral RNA by type III CRISPR system enhances diagnostic. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-1466718/v1 (2022).

49. Leenay, R. T. & Beisel, C. L. Deciphering, communicating, and engineering the CRISPR PAM. *J. Mol. Biol.* **429**, 177–191 (2017).

50. Silas, S. et al. Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. *eLife* **6**, e27601 (2017).

51. Estrella, M. A., Kuo, F.-T. & Bailey, S. RNA-activated DNA cleavage by the type III-B CRISPR-Cas effector complex. *Genes Dev.* **30**, 460–470 (2016).

52. Liu, L. et al. The molecular architecture for RNA-guided RNA cleavage by Cas13a. *Cell* **170**, 714–726 (2017).

53. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol.*

Biol. Evol. **30**, 772–780 (2013).

54. Steenwyk, J. L., Buida, T. J. 3rd, Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).

55. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

56. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

57. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).

58. Guzman, L. M., Belin, D., Carson, M. J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose P_{BAD} promoter. *J. Bacteriol.* **177**, 4121–4130 (1995).

59. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).

60. Lapinaite, A. et al. DNA capture by a CRISPR-Cas9-guided adenine base editor. *Science* **369**, 566–571 (2020).

61. Harris, C. J., Molnar, A., Müller, S. Y. & Baulcombe, D. C. FDF-PAGE: a powerful technique revealing previously undetected small RNAs sequestered by complementary transcripts. *Nucleic Acids Res.* **43**, 7590–7599 (2015).

62. Li, Z., Zhang, H., Xiao, R., Han, R. & Chang, L. Cryo-EM structure of the RNA-guided ribonuclease Cas12g. *Nat. Chem. Biol.* **17**, 387–393 (2021).

63. Bushnell, B., Rood, J. & Singer, E. BBMerge—accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017).

64. Green, M. R. & Sambrook, J. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 2012).

65. Leenay, R. T. et al. Identifying and visualizing functional PAM diversity across CRISPR-Cas systems. *Mol. Cell* **62**, 137–147 (2016).

66. Wandera, K. G. et al. An enhanced assay to characterize anti-CRISPR proteins using a cell-free transcription-translation system. *Methods* **172**, 42–50 (2020).

Acknowledgements

We thank K. Makarova for guidance on naming Cas12a2 and for feedback on phylogenetic analyses; L. Fläxl for assistance with transcription–translation; L. Ostertag for assistance with nick translation; F. Ttofali for assistance with plasmid cloning and testing; D. Collias for suggestions around nicking translation; and A. Pawluk and A. Özcan for feedback on the manuscript. The plasmids pET28a-mH6-Cas12g1 (Addgene plasmid, 120879) and pACYC-Cas12g1 (Addgene plasmid, 120880) were gifts from Arbor Biosciences. Plasmid pC001 was a gift from F. Zhang (Addgene plasmid, 79150). This work was supported by an ERC Consolidator grant (865973 to C.L.B.), the DARPA Safe Genes program (HR0011-17-2-0042 to C.L.B.), the National Institutes of Health (R35GM138080 to R.N.J.), the Netherlands Organization for Scientific Research (NWO) through a Rubicon Grant (project 019.193EN.032, to I.M.), and the Federal Agency for Disruptive Innovation (to C.L.B.). The views, opinions and/or findings expressed should not be interpreted as representing the official views or policies of the Department of Defense or the US Government.

Funding

Open access funding provided by Helmholtz-Zentrum für Infektionsforschung GmbH (HZI).

Author information

Author notes

1. Benjamin N. Gray

   Present address: Syngenta, Research Triangle Park, NC, USA

Authors and Affiliations

1. Helmholtz Institute for RNA-based Infection Research, Helmholtz Centre for Infection Research, Würzburg, Germany

   Oleg Dmytrenko, Elena Vialetto, Ioannis Mougiakos, Katharina G. Wandera, Johannes Weber, Thomas Gaudin & Chase L. Beisel

2. Benson Hill, St Louis, MO, USA

   Gina C. Neumann, Benjamin N. Gray & Matthew B. Begemann

3. Department of Chemistry and Biochemistry, Utah State University, Logan, UT, USA

   Thomson Hallmark, Dylan J. Keiser, Valerie M. Crowley, Hannah Domgaard, Josie Metcalf & Ryan N. Jackson

4. Medical Faculty, University of Würzburg, Würzburg, Germany

   Chase L. Beisel

Contributions

Conceptualization: O.D., G.C.N., R.N.J. and C.L.B. Methodology: O.D., G.C.N., V.M.C., J.M., H.D., T.H. and D.J.K. Bioinformatic discovery of Cas12a2 orthologues: B.N.G. and O.D. First observation of bacterial abortive infection: G.C.N. Phylogenetic analysis: O.D. Discovery of RNA targeting: O.D., D.J.K., R.N.J. and C.L.B. Experimentation in *E. coli*: O.D., G.C.N., V.M.C., E.V., I.M. and J.W. Experimentation in transcription–

translation: K.G.W., J.W. and O.D. Experimentation in vitro: V.M.C., D.J.K., H.D, T.H. and J.M. Nick-translation diagnostic: T.G. Writing—original draft: O.D., V.M.C., R.N.J. and C.L.B. Writing—review and editing: all of the authors. Visualization: O.D., D.J.K., T.H., H.D., R.N.J. and C.L.B. Supervision: M.B.B., R.N.J. and C.L.B. Funding acquisition: R.N.J. and C.L.B.

Corresponding authors

Correspondence to Matthew B. Begemann, Ryan N. Jackson or Chase L. Beisel.

Ethics declarations

Competing interests

Benson Hill has one granted patent (US 9,896,696) and has filed additional patent applications. G.C.N. and M.B.B. are employees of Benson Hill. O.D., R.N.J. and C.L.B. have filed provisional patent applications on the related concepts, with one granted patent (US 9,896,696). C.L.B. is a co-founder of Locus Biosciences and is a scientific advisory board member of Benson Hill. The other authors declare no competing interests.

Peer review

Peer review information

Nature thanks Raymond Staals and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Maximum-likelihood phylogeny of Cas12a2 and Cas12a nucleases, with Cas12b as an outgroup.

The phylogenetic tree generated using amino acid sequences is the annotated version of Fig. 1a. The shaded regions designate Cas12a2 (red), Cas12a (blue), and Cas12b (grey) orthologues. The blue and red filled circles indicate systems that contain both Cas12a2 and Cas12a. Transfer bootstrap expectation (TBE) values are shown at the nodes.

Extended Data Fig. 2 Cas12a2 can utilize Cas12a crRNA repeats and requires an RNA target in vivo.

a, Different guide:target pairs tested under antibiotic selection of the nuclease and target plasmids or only the nuclease plasmid. **b**, Reduction in plasmid transformation by Cas12a2 orthologues tested under antibiotic selection of the nuclease and target plasmids. **c**, Effect of swapping direct repeats associated with *Su*Cas12a2 and Cas12a nucleases on plasmid transformation. The indicated nuclease, repeat-encoding crRNA, and target were subjected to the traditional (left) and modified (right) plasmid interference assay in *E. coli*. **d**, Diagram of the target and non-target plasmids used in the plasmid clearance assay shown in e. **e**, Impact of a promoter and a terminator upstream of the target site on plasmid clearance by *Su*Cas12s2 and *Lb*Cas12a. **f**, Diagram of the cell-free TXTL reactions used to evaluate the impact of target expression on Cas12a2 collateral silencing shown in g. **g**, The effect of collateral silencing by *Su*Cas12a2 in TXTL as a function of having a promoter, no promoter, or terminator upstream of the target expression site. Scatter plots represent averages of 4 technical replicates ± s.d. Unless stated otherwise, values are means ± s.d. of at least 3 independent experiments started from separate colonies. ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.005$ calculated with one-tailed Welch's t-test.

Extended Data Fig. 3 Pre-crRNA processing by *Su*Cas12a2.

a, Diagram of the predicted secondary structure of the *Su*Cas12a2 direct repeat. Cleavage sites of *Su*Cas12a2 and Cas12a are indicated. **b**, Sequencing coverage of the cDNA mapped to the crRNA locus in *E. coli* BL21(AI) expressing *Su*Cas12a2. Coverage for all of the quality-filtered reads above 10 nts as well as reads between 10 and 55 nts mapped to the plus strand are shown. **c**, In vitro processing of a 56-nt pre-crRNA incubated for 60 min in the presence of apo-*Su*Cas12a2 or two mutants predicted to disrupt crRNA processing. **d**, In vitro *Su*Cas12a2-mediated cleavage of a pre-crRNA containing three direct repeats and three spacers (3× crRNA). Reactions containing apo-*Su*Cas12a2 are indicated. Time points after mixing the cleavage reaction with apo-*Su*Cas12a2 are indicated. The estimated sizes of the pre-crRNA and crRNAs after cleavage are indicated on the left. The last two lanes contain RNA extracted from *Su*Cas12a2 co-expressed with a CRISPR array. The difference in size between the major crRNA band in the in vitro assay (~ 67 nt) and the crRNA extracted from *Su*Cas12a2 bound to *E. coli*-expressed crRNA (~42) may be due to further trimming by 3′ exonucleases in the cell. For gel source data, see Supplementary Fig. 1.

Extended Data Fig. 4 Properties of target recognition and collateral cleavage by *Su*Cas12a2 .

a, Non-specific collateral activities of *Su*Cas12a2, *Lb*Cas12a, *Lwa*Cas13a, and *Ab*Cas12g towards FAM-labelled non-target ssDNA, dsDNA, and RNA in the presence of target ssDNA, dsDNA, and RNA. All cleavage reactions were conducted at 37 °C unless specified otherwise. *Ab*Cas12g was triggered by RNA and ssDNA and exhibited preferential collateral cleavage of RNA over ssDNA but no discernable cleavage of dsDNA. **b**, *Su*Cas12a2-mediated cleavage over time of FAM-labelled RNA (top), ssDNA (middle), and dsDNA (bottom) non-target substrates. These substrates are cleaved by *Su*Cas12a2 through its non-specific collateral activity. **c**, Electromobility shift assay indicating *Su*Cas12a2 and *Lb*Cas13a mostly do not indiscriminately degrade dsRNA. Uncleaved substrates and cleaved products are indicated. **d**, Impact of mutating conserved cysteines within the predicted Zinc finger domain of *Su*Cas12a2 on RNA-triggered collateral activity. The mutated cysteines within *Su*Cas12a2 are C1170S, C1173S, C1188S and C1191S. **e**, Electromobility shift assays of supercoiled pUC19 plasmid with *Su*Cas12a2:crRNA and different RNA sequences over time. RNA sequences

include: Non-self target (complementary to gRNA with 3′ GAAAG PFS), Target complement (complement of non-self target), Non-target (RNA sequence used in *trans*-cleavage assays), No PFS (RNA that only contains compliment to gRNA no 3′ PFS), Self-PFS (complementary to gRNA with 3′ AUCUA PFS). RNA sequences can be found in Supplementary Table 1. *Su*Cas12a2(E1063A) with the non-self target is included as a negative control. NN: no-nuclease control. For gel source data, see Supplementary Fig. 1.

Extended Data Fig. 5 PFS depletion screen with *Su*Cas12a2 in *E. coli* BL21(AI).

a, Sequencing coverage of the PFS libraries from the target and the control *E. coli* cultures. Data from two biological replicates are shown. **b**, Correlation between the depletion scores obtained from the two replicate libraries. **c**, The complex PFS profile recognized by *Su*Cas12a2. See ref. [65] for more information on interpreting PAM wheels. Given the complexity of the PFS profile, four different PAM wheels are shown based on each nucleotide at the −1 PFS position. **d**, Validation of selected PFS sequences identified in the PFS screen with *Su*Cas12a2 using the plasmid clearance assay. Values are means ± s.d. of ≥ 3 independent experiments started from separate colonies. ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.005$ calculated with one-tailed Welch's t-test.

Extended Data Fig. 6 Impact of *Su*Cas12a2 guide and PFS alterations on plasmid targeting in *E. coli* BL21(AI).

a, Impact of guide length on plasmid clearance. The guide sequence is depicted with blue letters. The standard crRNA guide length based on crRNA processing (Extended Data Fig. 3) is 24 nts. **b**, Reduction in plasmid transformation by *Su*Cas12a2 and *Lb*Cas12a in the presence of a shared PAM/PFS, Cas12a2-specific PFS, and mismatches to the target under target plasmid and nuclease plasmid or nuclease plasmid-only selection. Values are means ± s.d. of at least 3 independent experiments started from separate colonies. ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.005$ calculated with one-tailed Welch's t-test.

Extended Data Fig. 7 *Su*Cas12a2 degrades cellular RNA.

a, 1% agarose gel resolving total RNA extracted from *E. coli* BL21(AI) expressing *Su*Cas12a2, *Lb*Cas12a and *Ls*Cas13a under target (T) and non-target (NT) conditions. Expression of the nucleases and crRNA was induced with 10 nM IPTG and 0.2% arabinose for 2 h. Individual wells for each condition represent biological replicates. Nucleotide sizes are based on a resolved RNA ladder. sRNA pool: RNA pool comprising tRNAs and other small RNAs. **b**, Quantification of band intensities across the 4 independent experimental replicates started from separate colonies. Significance was calculated using two-tailed Welch's t-test (ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.005$).

Extended Data Fig. 8 Flow cytometry analysis of the nuclease-expressing *E. coli* cells without antibiotic selection.

The cultures were collected 4 h after inoculation and induction of nuclease and crRNA. Prior to the analysis, the cells were stained with DAPI. The subpopulation with low DAPI and high forward scatter reflects elongated cells with reduced DNA content. Contour plots are representative of 4 independent experiments.

Extended Data Fig. 9 Confocal laser scanning micrographs of *E. coli* BL21(AI) transformed with *Su*Cas12a2–crRNA and *Lb*Cas12-crRNA expression plasmids under target and non-target conditions in the absence of antibiotic selection.

Extensive filamentation can be observed with the bacteria containing the target-expression plasmid and expressing *Su*Cas12a2 and target crRNA 2 h and 4 h after induction. Within each overlaid image, fluorescent DNA (DAPI) staining is shown in blue and membrane (FM4–64) staining is shown in red. Paired images are representative of 2 independent experiments.

Extended Data Fig. 10 Assessment and extension of Cas12a2-based diagnostics.

a, Progress curve for target RNA-activated cleavage of RNA beacon by *Su*Cas12a2 at 37 °C. **b**, Limit of detection of *Su*Cas12a2 using an RNA beacon as determined using the velocity method[46]. Velocities were obtained by regression analysis of the linear regions of the progress curves. The velocity method was used to determine all reported limits of detection. **c**, Progress curve for target RNA-activated cleavage of ssDNA beacon by Cas12a2 at ambient temperature (RT). Error bars in a and c represent the mean and standard deviation of three independent measurements. Error bars in b represent the standard error for the linear fit of three independent measurements. **d**, Overview of detection assay based on plasmid nicking and nick translation. **e**, Resolved plasmid substrate incubated with the ca33Cas12a2-gRNA RNP (100 nM) and the indicated RNA (1 nM). The incubation time before initiating nick translation is indicated for each experimental condition. The first lane (Plasmid only) represents the plasmid incubated without an RNP. Non-target RNA1: *COA1* gRNA and SARS-CoV-2 RNA. Non-target-RNA2: SARS-CoV-2 gRNA and COA1 RNA. **f**, Fluorescence measurements following nick translation with the different reactions from e. Fluorescence values were normalized to that of the plasmid-only control. For the positive control, the plasmid was subjected to a mix of DNase I and DNA polymerase I. Values are means ± s.d. of 3 independent experiments. Significance was calculated using two-tailed Welch's t-test compared to a value of 1 (ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.005$). For gel source data, see Supplementary Fig. 1.

Supplementary information

Supplementary Fig. 1

Gel source data.

Reporting Summary

Peer Review File

Supplementary Figures

Supplementary Figs. 2–9 and the legend for Supplementary Fig. 1.

Supplementary Tables

Supplementary Tables 1–5.

Supplementary File 1

Fasta file listing Cas12a2, Cas12a and Cas12b amino acid sequences.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Dmytrenko, O., Neumann, G.C., Hallmark, T. *et al.* Cas12a2 elicits abortive infection through RNA-triggered destruction of dsDNA. *Nature* **613**, 588–594 (2023). https://doi.org/10.1038/s41586-022-05559-3

- Received: 13 June 2022

- Accepted: 11 November 2022
- Published: 04 January 2023
- Issue Date: 19 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05559-3

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

This article is cited by

- **RNA targeting unleashes indiscriminate nuclease activity of CRISPR–Cas12a2**
  - Jack P. K. Bravo
  - Thomson Hallmark
  - David W. Taylor

  Nature (2023)

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05559-3